Volume 92, Number 1, February 1984

Zvi Eckstein: A Rational Expectations Model of Agricultu

Jeremy I. Bulow and Lawrence H Summers The Ta
Risky Assets

Julio J. Rotemberg: A Monetary Equilibrium Mod
Transactions Costs

Martin D. Dooley and Peter Gottschalk Earnings Ineq
Males in the United States: Trends and the E
Labor Force Growth

Robert P. Flood and Peter M Garber Gold Monetization
Gold Discipline

Boyan Jovanovic: Matching, Turnover, and Unemployment

Bennett T McCallum: Are Bond-financed Deficits Inflationary?
A Ricardian Analysis

Russell D. Roberts: A Positive Model of Private Charity and
Public Transfers

The University of Chicago Press

JOURNAL OF
POLITICAL ECONOMY

Edited by
ROBERT J. BARRO SAM PELTZMAN*
JACOB A. FRENKEL* JOSÉ A. SCHEINKMAN
JAMES J. HECKMAN GEORGE J. STIGLER

In cooperation with OTHER MEMBERS of the DEPARTMENT OF
ECONOMICS and the GRADUATE SCHOOL OF BUSINESS
of the UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES

Editorial Assistants: VICKY M. LONGAWA and ANN LINDNER

*On leave.

The Journal of Political Economy (ISSN 0022-3808) is published bimonthly in February, April, June, August, October, and December by the University of Chicago Press. Subscription rates, U.S.A.: institutions, 1 year $50.00; individuals, 1 year $29.00. Student subscription rate, U.S.A.: 1 year $18.00 (letter from professor must accompany subscription). Other countries add $4.50 for each year's subscription to cover postage. Single copy rates: *institutions $8.50, individuals $5.00. Back issues are available from 1978 (vol. 86).* Make all remittances payable to *Journal of Political Economy,* The University of Chicago Press, in United States currency or its equivalent. **Business correspondence** should be addressed to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, Illinois 60637.

Claims for missing numbers should be made within the month following the regular month of publication. The publishers expect to supply missing numbers free only when losses have been sustained in transit and when the reserve stock will permit.

Letters to the editors and manuscripts should be addressed to the Editor of the *Journal of Political Economy,* 1126 East 59th Street, Chicago, Illinois 60637. **Manuscripts should be submitted in triplicate, accompanied by a $40.00 submission fee made payable to the Journal.** The proceeds from the submission fees are used to pay for refereeing services. Accepted manuscripts must be typed according to the University of Chicago *Manual of Style.* References should be typed double-spaced at the end of the article. Footnotes should be numbered in sequence and double-spaced following the references. Tables should follow the footnotes. Originals of the figures, drawn in india ink, should be submitted if the manuscript is accepted. Abstracts not exceeding 100 words should be submitted in duplicate along with the manuscript.

Copying beyond Fair Use: The code on the first page of an article in this journal indicates the copyright owner's consent that copies of the article may be made beyond those permitted by Sections 107 or 108 of the U.S. Copyright Law provided that copies are made only for personal or internal use, or for the personal or internal use of specific clients and provided that the copier pay the stated per-copy fee through the Copyright Clearance Center, Inc. Operation Center, P.O. Box 765, Schenectady, New York 12301. To request permission for other kinds of copying, such as copying for general distribution, for advertising or promotional purposes, for creating new collective works, or for resale, kindly write to the publisher.

Reprinted volumes 1–85 available from Johnson Reprint Corp., 111 Fifth Avenue, New York, New York 10003. Volumes available **in microfilm** from University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106; **in microfiche** from KTO Microform, Route 100, Millwood, New York 10546.

Notice to subscribers: If you change your address, please notify us and your local postmaster immediately, giving *both* your old and your new address. *Allow four weeks for the change.* **Postmaster:** Send address changes to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, Illinois 60637.

Second-class postage paid at Chicago, Illinois, and at additional mailing office.

© 1984 by The University of Chicago. All rights reserved.

Journal of Political Economy

Volume 92 Number 1 February 1984

Articles

1 A Rational Expectations Model of Agricultural Supply
Zvi Eckstein

20 The Taxation of Risky Assets
Jeremy I. Bulow and Lawrence H. Summers

40 A Monetary Equilibrium Model with Transactions Costs
Julio J. Rotemberg

59 Earnings Inequality among Males in the United States: Trends and the Effect of Labor Force Growth
Martin D. Dooley and Peter Gottschalk

90 Gold Monetization and Gold Discipline
Robert P. Flood and Peter M. Garber

108 Matching, Turnover, and Unemployment
Boyan Jovanovic

123 Are Bond-financed Deficits Inflationary? A Ricardian
Bennett T. McCallum

136 A Positive Model of Private Charity and Public Transfers
Russell D. Roberts

Comments

149 Misconceptions about the Real-Bills Doctrine: A Comment on Sargent and Wallace
David Laidler

156 Inferior Forecasters, Cycles, and the Efficient-Markets Hypothesis: A Comment
Hersh M. Shefrin

162 Efficient Asset Portfolios and the Theory of Normal Backwardation: A Comment
Alan J. Marcus

Miscellany

165 Financial Futures Markets and Tabular Standards
Milton Friedman

Book Reviews

168 Helmut Frisch (ed.), *Schumpeterian Economics*
Scott Gordon

170 Alec Nove, *The Economics of Feasible Socialism*
Nicolas Spulber

Journal of Political Economy

Volume 92 Number 2 April 1984

Articles

175 The Trade-Off between Equality and Efficiency
Edgar K. Browning and William R. Johnson

204 The Effects on Output of Money Growth and Interest Rate Volatility in the United States
Paul Evans

223 Crime on the Court
Robert E. McCormick and Robert D. Tollison

236 The Effects of Increased Copyright Protection: An Analytic Approach
Ian E. Novos and Michael Waldman

247 Conversion to a Consumption Tax: The Transition in a Life-Cycle Growth Model
Laurence S. Seidman

268 Aggregate Information and the Role of Monetary P
Open Economy
Kent P. Kimbrough

286 International Capital Movements under Uncertainty
Gene M. Grossman and Assaf Razin

307 Are Economic Time Series Asymmetric over the Business Cycle?
Salih N. Neftçi

Confirmations and Contradictions

329 The Incidence of a Tax on Pure Rent: The Old Reason for the Old Answer
George Fane

Book Reviews

334 Michael Edelstein, *Overseas Investment in the Age of High Imperialism: The United Kingdom, 1850–1914*
C. Knick Harley

337 Frank Hahn, *Money and Inflation*
Herschel I. Grossman

340 Robert E. Hall and Alvin Rabushka, *Low Tax, Simple Tax, Flat Tax*
Joseph A. Pechman

343 Keith S. Rosenn, *Law and Inflation*
Meir Heth

345 T. Paul Schultz, *Economics of Population*
Boone A. Turchi

Journal of Political Economy

Volume 92 Number 3 June 1984

Articles

349 The Informational Efficiency of Experimental Asset Markets
Daniel Friedman, Glenn W. Harrison, and Jon W. Salmon

409 Futures Markets and Production Decisions
Alan J. Marcus and David M. Modest

427 Advertising as a Signal
Richard E. Kihlstrom and Michael H. Riordan

451 A Simple Theory of International Trade with Multinational Corporations
Elhanan Helpman

472 Restrictions on Price Advertising
Michael Peters

486 Work Incentives, Hierarchy, and Internal Labor Ma
James M. Malcomson

508 Sraffa, Labor Theories of Value, and the Economi
Wage Rate Determination
Edwin Burmeister

527 Bequests and the Size of Population When Population Is Endogenous
Marc Nerlove, Assaf Razin, and Efraim Sadka

Comments

532 The Decline in Male Labor Force Participation: Comment
Robert H. Haveman and Barbara L. Wolfe

542 Disability Insurance and Male Labor Force Participation: A Response to Haveman and Wolfe
Donald O. Parsons

Book Reviews

550 Jagdish N. Bhagwati, ed., *Import Competition and Response*
Rachel McCulloch

553 Robert E. Hall, ed., *Inflation: Causes and Effects*
Jeremy Greenwood

558 Dennis C. Mueller, ed., *The Political Economy of Growth*
Anne O. Krueger

560 Michael Sattinger, *Capital and the Distribution of Labor Earnings*
Alan Harrison

Journal of Political Economy

Volume 92 Number 4 August 1984

565 In Memoriam: Harry Gordon Johnson (1923–1977)

567 Harry Johnson's Contributions to International Trade Theory
W. M. Corden

592 Harry Johnson as a Macroeconomist
David Laidler

616 Harry G. Johnson as a Development Economist
Arnold C. Harberger and David Wall

642 Harry Johnson as a Social Scientist
Richard E. Caves

659 Harry G. Johnson: A Bibliography
Vicky M. Longawa, Editor and Compiler

Articles

712 Schooling, Search, and Spouse Selection: Testing [illegible] Theories of Marriage and Household Behavior
Bryan L. Boulier and Mark R. Rosenzweig

733 Entry, Industry Growth, and the Microdynamics of [illegible] Supply
John C. Hause and Gunnar Du Rietz

758 Job Matching with Finite Horizon and Risk Aversion
Milton Harris and Yoram Weiss

Book Reviews

780 Maurice Allais, *La Théorie générale des surplus*
Roger Guesnerie

781 Victoria Chick, *Macroeconomics after Keynes: A Reconsideration of the* General Theory
Robert M. Solow

787 Victor R. Fuchs, *How We Live: An Economic Perspective on Americans from Birth to Death*
Lee Benham

Journal of Political Economy

Volume 92 Number 5 October 1984

Articles

791 Violations of the Gold Points, 1890–1908
Truman A. Clark

824 An Equilibrium Model of Search Unemployment
James W. Albrecht and Bo Axell

841 The Effect of the Discount Rate on Depletion of Exhaustible Resources
Y. Hossein Farzin

852 An Estimable Dynamic Stochastic Model of Fertility and Child Mortality
Kenneth I. Wolpin

875 Cross-Regime Evidence of Macroeconomic Rationality
Roger C. Kormendi and Philip G. Meguire

909 On Success and Dissolution of the Labor-managed F
Capitalist Economy
Hajime Miyazaki

932 Regulation and Industrial Organization
Andrew F. Daughety

951 On Measuring Natural Resource Scarcity
Robert Halvorsen and Tim R. Smith

965 New Products and the Factor Content of International Trade
Richard A. Brecher and Ehsan U. Choudhri

972 Angell and the Stable Money Rule
Joong-Koon Lee and Donald C. Wellington

Comment

979 Aggregation and the Factoral Content of Trade
Richard Gift and William Marxsen

Book Reviews

985 Joseph Friedman and Daniel H. Weinberg, *The Economics of Housing Vouchers*
John C. Weicher

988 T. R. Nicholson, *The Birth of the British Motor Car, 1769–1897* (3 vols.)
Nathan Rosenberg

Journal of Political Economy

Volume 92 Number 6 December 1984

Articles

991 The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory
Alvin E. Roth

1017 On Interindustry Differences in Absolute Productivity
William J. Baumol and Edward N. Wolff

1035 Price Leadership and Dynamic Aspects of Oligopoly in U.S. Manufacturing
Micha Gisser

1049 The New Divisia Monetary Aggregates
William A. Barnett, Edward K. Offenbacher, and Paul A. Spindt

1086 Job Matching and Occupational Choice
Robert A. Miller

1121 Tenancy Choice in a Competitive Framework with [illegible]tions Costs
Lee J. Alston, Samar K. Datta, and Jeffrey B. Nugent

1134 Unionization and Profitability: Evidence from the [illegible] Market
Richard S. Ruback and Martin B. Zimmerman

Comments

1158 A Note on Extraction with Nonconvex Costs
Sheldon Kimmel

1168 Competitive Equilibria in Exhaustible Resource Markets with Decreasing Costs: A Comment on Eswaran, Lewis, and Heaps's Demonstration of Nonexistence
Gene E. Mumy

Book Reviews

1175 Herman H. Gossen, *The Laws of Human Relations*
George J. Stigler

1176 Joseph P. Kalt, *The Economics and Politics of Oil Price Regulation: Federal Policy in the Post Embargo Era*
Henry D. Jacoby

1178 Réal P. Lavergne, *The Political Economy of U.S. Tariffs: An Empirical Analysis*
Jonathan J. Pincus

1183 Index to Volume 92

A Rational Expectations Model of Agricultural Supply

Zvi Eckstein

Yale University and Tel-Aviv University

This work suggests a dynamic linear rational expectations modeling strategy for using time-series observations to study the impact of product prices on agricultural production and land allocation. The farmer's decision consists of specifying each period the portion of his land to be devoted to the growing of alternative crops. The dynamic element in the farmer's problem comes from the technology. A quadratic production technology is used. The farmer's optimal land allocation decision rules are derived as linear functions of past land allocations, expectations of future product prices, and other exoge-nous variables; and they can be interpreted as an optimal crop [illegible] tion. Assuming rational expectations and the exogeneity of p[illegible] for a small open economy, one obtains closed-form linear regr[illegible] equations representing decision rules and stochastic proce[illegible] These are then estimated and tested using maximum likeli[illegible] methods and aggregate data from the Egyptian agriculture

I. Introduction

Using annual data, economists have suggested different theoretical and empirical methods to evaluate farmers' responses to changes in crop prices. The existence of persistent patterns of serial and cross correlations between land allocations, production, and prices has been observed and debated in the economic literature for many years.

This paper is based on my 1981 Ph.D. dissertation at the University of Minnesota. I am grateful to Anne O. Krueger, Lars Hansen, Thomas Sargent, Christopher Sims, Vernon Ruttan, Rao Aiyagari, Martin Eichenbaum, Wallace Huffman, Dan Peled, T. N. Srinivasan, and Ken Wolpin for useful comments on previous drafts of this paper. Comments by the referees and editor of this journal were most useful.

[*Journal of Political Economy*, 1984, vol. 92, no. 1]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0001$01.50

The fact that output selling price is not known when input decisions are made and the resulting need for farmers to make decisions based on expectations of future prices have been suggested as the main reasons for the cyclical movements of output. Early single-equation estimates, with current output as a function of only one past price, showed small linkage between prices and output. Later, Nerlove (1956, 1958) showed that the static supply equation, with the assumptions of "partial adjustments" in actual output (land) and adaptive expectations of the output price, gave rise to a single distributed lag model that could explain much of the supply response to output price changes. Askari and Cummings (1976) report more than 600 estimates of different versions of Nerlove's model for many crops and countries. Muth (1961) criticized Nerlove's model in the light of his rational expectations model.[1]

In this study, an empirical model of agricultural supply is derived from a dynamic and stochastic framework where farmers are assumed to maximize the expected present value of profit by choice of land allocations subject to a dynamic and stochastic technology as well as uncertain price movements.[2] The analysis focuses on the dynamics of crop areas and their joint movement with crop prices. In this context, it is straightforward to show that rational farmers are unlikely to interpret price fluctuations that are serially uncorrelated as signaling a permanent alteration in the incentives confronting them. Furthermore, any permanent or transitory change in taxes, subsidies, or tariffs affects the dynamic response of the cropped area, such that the structural form of the land allocation equation varies directly with policy rules. Consequently, predictions with respect to changes in policy require complete identification of the structural parameters that govern the production and the price processes (Marschak 1953; Lucas 1976).

Understanding the consequences of price movements on agricultural production is of particular importance in less developed countries. Therefore, the model is implemented using aggregate agricultural data from Egypt, 1913–69. The dynamic statistical properties of the data were initially analyzed by the estimation of vector autoregressions (VAR) that included cotton and wheat cropped areas, yields, and prices. The results showed (1) that the percentage of forecast

---

[1] See Behrman (1968) for a detailed discussion of the issues and complete country work that follows the Nerlovian model. More recently, Nerlove (1979) critically reviewed the traditional supply response model in the light of recent developments in economic time-series models, e.g., Muth (1961); Lucas (1976), and Nerlove, Grether and Carvalho (1979). In my dissertation (Eckstein 1981*b*) there is some further discussion of the subject.

[2] This approach follows Sargent (1979*b*, 1981) and is consistent with Schultz's (1978, p. 4) view on farmers' behavior.

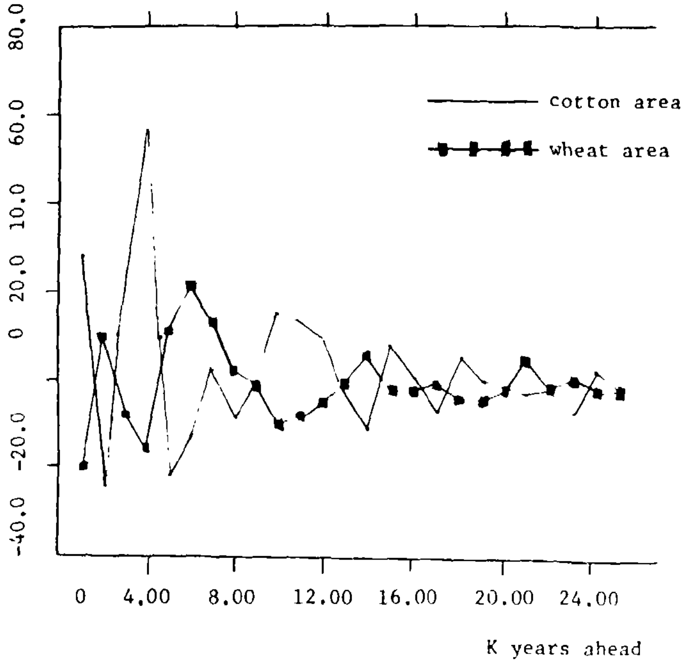


FIG. 1 —Responses of cotton and wheat areas to a shock in cotton price over price

error produced by a shock to prices accounted for 15–50 perc the total variance in land allocations and yields. Hence, price ments have a significant effect on farmers' decisions. (2) A sho any variable causes cotton and wheat areas to respond in oppo ways and to fluctuate frequently as they converge to their mean. Figure 1 shows this result for an innovation in the cotton price over the wheat price in a particular system.[3] It is of interest now to find out whether a rational expectations model that accounts for some technological aspects of agricultural production can interpret well the properties of the Egyptian data.

The remainder of the paper is organized as follows: In Section II, I construct a model where the dynamics in the production process are capable of producing a dynamic land allocation decision rule that exhibits the "cobweb phenomenon"—frequent fluctuations in crop areas, depicted by figure 1. In Section III, I present results from *maximum likelihood estimation and tests of the model.* Concluding remarks are contained in the last section.

---

[3] The methodology for estimating and interpreting VAR models was developed by Sims (1980), who used it to analyze macroeconomic questions. T. Doan and R. Litterman's package of Regression Analysis of Time Series (RATS), version 4.1, has been used for computations. The results from estimating several unrestricted VARs have preceded the formulation of the models. I estimated several different VARs and the results turned out almost the same. Detailed information exists in my dissertation (Eckstein 1981*b*). Figure 1 is taken from a system of cotton and wheat areas and yields and the relative price (cotton price over wheat price).

II. A Dynamic Land Allocation Model

In this section I analyze the stochastic dynamic optimization problem of a farmer endowed with land that can be allocated between two crops (e.g., cotton and wheat). The model shows that if the cultivation of at least one crop (e.g., cotton) results in deterioration of land productivity, the optimization yields a dynamic land allocation process that can be estimated using aggregate time-series data.

The role of adjustment costs in affecting input decisions and output supply has been analyzed extensively in the economics literature (e.g., Lucas 1967; Gould 1968; Sargent 1979*b*) and to some extent also in the agricultural economics literature (e.g., Nerlove 1958, 1972; Nerlove et al. 1979). Adjustment costs in annual crop production could be justified when land preparation of a plot has crop-specific requirements. When land is continuously cultivated, the issue of substitution and complementarity effects in the production of alternative crops becomes important. The depletion of nitrogen from the soil is an important direct constraint on the development of land fertility and the production of all crops. Furthermore, monoculture causes an accumulation of crop-specific insects and worms, which have an important indirect effect on the actual crop yield from the land. Hence, the current productivity of land for a given crop depends on the cropping history of a plot of land.[1]

To develop a simple model that contains these technological characteristics, consider the following definitions:

X_{it} is the production of crop i at time t;
P_{it} is the price that farmers receive for the production of crop i at time t;

[1] Crop rotation is the well-known method to prevent the direct and the indirect deterioration in land productivity under continuous cultivation. Fertilizer and pesticides are the main inputs that directly control land productivity by building up the content of the soil and eliminating insects and worms. For the importance of crop rotation and land fertility deterioration in the United States, the reader is referred to Nerlove (1958, pp. 122–35) and the many publications on this subject by the U.S. Department of Agriculture. In Egypt, the case study here, Hansen and Marzouk (1965) and Hansen and Nashashibi (1975) emphasized the crucial effect of nitrate depletion in cotton production on the acreage allocation of all crops.

A_{it} is the land allocated at time $t - 1$ for the production of crop i at time t;
$\bar{A}$ is the total available cultivated land at time t;
$0 < \beta < 1$ is the objective discount factor;
a_{it} is the shock to production of crop i at time t;
E is the mathematical expectation operator, where $E_t(X) = E(X|\Omega_t)$ and Ω_t is the information set available for farmers at time t; and
L is the lag operator defined by the property $L^kX_t = X_{t-k}$.

Suppose that only crop 1 (cotton) is subject to dynamic constraints in production; therefore I can write

$$X_{1t} = F^1(A_{1t}, a_{1t}, A_{1t-1}, A_{1t-2}, \ldots; \mathbf{K}, \bar{A}), \tag{1}$$

where $F^1_1 > 0$, $F^1_2 > 0$, $F^1_j \leq 0$ for all $j \geq 3$, $F^1_{11} \leq 0$, $\mathbf{K}$ is a vector of other (fixed) inputs, and $\bar{A}$ is the size of a given plot of cultivated land. The negative sign of the derivative of the production of crop 1 with respect to lagged allocations of land for the same crop (F^1_j, $j \geq 3$) is due to the assumption of deterioration in productivity with respect to an increase in the past intensity of use of the same plot with the same crop. The production function of crop 2 (wheat) is given simply by

$$X_{2t} = F^2(A_{2t}, a_{2t}; \mathbf{K}, \bar{A}), \tag{2}$$

where $F^2_1 > 0$, $F^2_2 > 0$, and $F^2_{11} \leq 0$.[5] The representative farmer is assumed to maximize his expected discounted profit in terms of price of crop 1 (cotton) by choosing a contingency plan at t for allocating his given area ($\bar{A}$) for the time $t + 1$ prod two crops. Hence, the farmer's objective is to maximize

$$E_{-1} \lim_{N \to \infty} \sum_{t=0}^{N} \beta^t \left(X_{1t} + \frac{P_{2t}}{P_{1t}} X_{2t} \right)$$

subject to the land constraint

$$A_{1t} + A_{2t} = \bar{A} \tag{3}$$

and the production functions (1) and (2).[6] The optimization is subject to the given levels of A_{1j}, $j < 0$, the well-defined joint distribution of

---

[5] The partial derivative of F^i ($i = 1, 2$) with respect to the variable in the jth position in the function is F^i_j. The asymmetry between cotton (1) and wheat (2) production technology is motivated by the Egyptian case, where cotton productivity is severely affected by past cultivation patterns but wheat production is not.

[6] The choice of the price in the numerator affects the solution to the optimization problem. Here I divide by P_{1t} in order to preserve the quadratic form of the model.

$\{(P_{2t}/P_{1t}), a_{1t}, a_{2t}\}_{t=0}^{\infty}$, and the assumed information set at the time the decision for period t is made; i.e.,

$$\Omega_{t-1} = \Big\{A_{1t-1}, A_{1t-2}, \ldots, a_{1t-1}, a_{1t-2}, \ldots, \frac{P_{2t-1}}{P_{1t-1}}, \frac{P_{2t-2}}{P_{1t-1}}, \ldots,$$
$$S_{t-1}, S_{t-2}, \ldots\Big\}.$$

Given that certain additional regularity conditions are met, the optimization problem (3) has a unique solution for the land allocation for crop 1 at time t that is a function of the information set Ω_{t-1}. Since this paper is concerned with the empirical implications of the model for aggregate data, I seek an analytical solution to problem (3). I suggest a linear-quadratic approximation for the production function above and the stochastic processes of the uncontrolled variables, and I assume that both crops can be produced on the same plot at the same period (4).[7] In particular, let (1) be specified as

$$X_{1t} = \left(f_1 + a_{1t} - \frac{g_1}{2}A_{1t}\right)A_{1t} + d_1\left(1 - \frac{A_{1t-1}}{\bar{A}} - \frac{A_{1t}}{\bar{A}}\right)A_{1t}, \quad (5)$$

where f_1, g_1, and d_1 are positive scalars. The first term represents a standard quadratic function. The second term is meant to approximate the deterioration in land productivity. For $d_1 > 0$, the particular approximation suggests that if the summation of the fractions of land from last and current periods is greater than one, then the current average productivity of land is reduced. Furthermore, if the summation of $A_{1t}/\bar{A}$ and $A_{1t-1}/\bar{A}$ is less than one, the current cultivation of crop 1 is assumed to be on land that has been used for crop 1 for only the current year. Hence, the average productivity is increased. If the sum of $A_{1t}/\bar{A}$ and $A_{1t-1}/\bar{A}$ is equal to one, there is no linkage between the current average productivity of land and past cultivations. Hence, only if it turns out that $A_{1t}/\bar{A} = ½$ for all $t \geq 0$ would the farmer's problem seem to be static. Observe that if $d_1 < 0$, the production function (5) implies that $F_3^1 > 0$; i.e., past cultivation with the same crop increases current productivity. This can be explained as arising from "high costs" of crop-specific plot preparations, which hold for several seasons.

---

[7] Alternatively one may assume that each plot is rigidly defined and that it can be used either for crop 1 or for crop 2. In this case the land allocation decision is a discrete choice problem, which is not tractable and is not an attractive description of the country aggregate or (even) the farm data. However, in the discrete case the production function (5) is intuitively very appealing, and if one has micro data on the history of plots one may use the above approximation to the farmer problem in order to derive a probit model.

The production function of crop 2 (wheat) is assumed to be linear:

$$X_{2t} = (f_2 + a_{2t})A_{2t}. \tag{6}$$

The specification of the model above is of particular interest since it enables one to (i) derive an analytical solution to the farmer's problem (3); (ii) emphasize the dynamic aspects of the production function of the main crop (cotton); (iii) keep the model close to the dynamic supply model of Nerlove (1958); and (iv) show what sort of restrictions one should impose on the theoretical farmer's problem in order to justify *explicitly* the linear time-series econometric models that are widely used. The production functions (5) and (6) are very simple approximations, but the Egyptian aggregate data did not a priori reject the specification of the cotton production function (5).[8] In general, an extended version of (5) that includes more lagged area and other inputs (e.g., labor and fertilizer) could be specified (see the App.), but only numerical solutions to the farmer's problem are feasible for much more complicated models.

If we substitute (4)–(6) into (3), the farmer's problem becomes: maximize

$$J = E_{-1} \lim_{N\to\infty} \sum_{t=0}^{N} \beta^t \Big[(f_1 + a_{1t})A_{1t} - \frac{g_1}{2} A_{1t}^2 + \frac{d_1}{\bar{A}} (\bar{A} - A_{1,t-1} - A_{1t})A_{1t} - R_t A_{1t} + R_t \bar{A} \Big], \tag{7}$$

by choice of $A_{10}, A_{11}, A_{12}, \ldots$, where $R_t = (1/P_{1t})\,[P_{2t}(f$
"real" shadow price for crop 1 land allocations. The op
subject to a given level of $A_{1,t-1}$ and a given law of m
stochastic processes of a_{1t}, R_t, and $\mathbf{S}_t$; i.e.,

$$\delta(L)\mathbf{Z}_t = \mathbf{U}_t,$$

where $\mathbf{Z}_t' = (a_{1t}, R_t, \mathbf{S}_t)$, and $\delta(L) = I - \delta_1 L - \delta_2 L^2 -$
where $\mathbf{S}_t$ is a vector of $n - 2$ exogenous variables at time t (suc
taxes, tariffs, and other variables that contain information on P_t's and a_1's) and δ_j is an $n \times n$ matrix for $j = 1, \ldots, k$. Further, $\mathbf{U}_t$ is an $n \times 1$ vector where $E(\mathbf{U}_t|\Omega_{t-1}) = 0$ and $E(\mathbf{U}_t\mathbf{U}_t') = \Sigma_t$ and where Σ_t is a positive semidefinite matrix. The vector stochastic process (8) is as-

---

[8] In Eckstein (1981*b*) I estimated it directly using aggregate data from Egypt, 1895–1975. (Data on prices are available only from 1913 and have been fixed by the government since 1968.) I divided (5) by A_{1t} to get a linear yield equation with a first-order serially correlated error (a_{1t}—see Sec. III). The results from the two-step efficient method of estimation were that (5) is significantly quadratic (i.e., $[(g_1/2) + (d_1/\bar{A})] > 0$), the serial correlation coefficient of a_{1t} is a significant positive fraction, and d_1 is an insignificant positive number. Furthermore, a likelihood ratio test for a structural change in (5) after World War II was rejected at the 5 percent significance level.

sumed to be of mean exponential order less than $1/\sqrt{\beta}$, so that a constant and a linear trend can be part of the vector $\mathbf{S}_t$. The variables in the vector $\mathbf{Z}_t$ $(n \times 1)$ are viewed as being unaffected by the farmer's decisions. Thus, for example, prices are assumed to be exogenously given to the representative farmer.

Following Sargent (1979*b*), the Euler equations of the dynamic problem (7) can be written as

$$A_{1t} = \lambda_1 A_{1t-1} - \frac{\lambda_1 \bar{A}}{d_1} \sum_{j=0}^{\infty} (\beta\lambda_1)^j [f_1 + d_1 + E_{t-1}(a_{1t+j}) - E_{t-1}(R_{t+j})] \qquad (9)$$

for all $t = 0, 1, 2, \ldots$, where λ_1 is a function of $g_1, d_1, \bar{A}$, and β, $-1 < \lambda_1 < 0$ if $d_1 > 0$, and $0 < \lambda_1 < 1$ if $d_1 < 0$.[9]

For any arbitrary set of expectations, (9) implies that

$$\frac{\partial A_{1t}}{\partial E_{t-1}(R_t)} < 0 \qquad (10a)$$

and

$$\frac{\partial A_{1t}}{\partial E_{t-1}(R_{t+1})} > 0, \qquad (10b)$$

for $d_1 > 0$. Equation (10a) is the standard result that more land is allocated to crop 1 as the expected price of the season goes up. However, (10b) implies that if farmers expect that in the following year the price of crop 2 relative to the price of crop 1 is going to decrease, they will decrease the quality of current land allocated to crop 1. This result could be viewed as a rationale for the "cobweb phenomenon," that is, that the inherent dynamics in the production function cause cotton and wheat production to oscillate in response to changes in the (rational) expected price in the future. Dynamic models with adjustment costs in land allocations $(d_1 < 0)$ imply the same result for (10a) but the opposite sign for (10b).

The assumption of rational expectations in this model implies that farmers maximize (7) subject to the true stochastic process of the exogenous variables (8). Using Hansen and Sargent's (1980) prediction method, we can write the optimal decision rule as a function of variables in the farmer's information set at time t; i.e.,[10]

---

[9] In Eckstein (1981*a*) I prove the equivalence of problem (7) where $d_1 < 0$ with the standard adjustment-costs problem discussed in the literature. The Nerlovian supply response model uses the adjustment-costs argument to justify adjustment in actual area vis-à-vis desired land allocations. If I include more than a 1-year deterioration effect in (5), the number of lags of land allocations in (9) will be equal to the number of years in the cumulative dynamic factor in the production function (see Hansen and Sargent 1981). If (5) is linear—i.e., $(g_1/2) + (d_1/A) = 0$—then the optimization problem (7) is not concave in A_1, and either there is a solution where $A_{1t} = A \leq \bar{A}$ for all t, or $A_{1t} = \bar{A}$ for some t and $A_{1t} = 0$ otherwise ("bang-bang"). In my (1981*b*) dissertation I show that with constant prices both cases are possible, depending on the prices of wheat and cotton.

$$A_{1t} = \lambda_1 A_{1t-1} + \gamma + \mu_1(L)a_{1t-1} + \mu_2(L)R_{t-1} + \mu_3(L)S_{1t-1} + \mu_4(L)S_{2t-1} + \dots + \mu_n(L)S_{n-2t-1}, \tag{11}$$

where

$$\gamma = \frac{\lambda_1 \bar{A}(f_1 + d_1)}{d_1\beta(1 - \beta\lambda_1)}$$

and

$$\mu_i(L) = \mu_{i0} + \mu_{i1}L + \dots + \mu_{ij}L^j$$

for all $i = 1, 2, \dots, n$ and $j \leq k$. Equation (11) is an exact closed-form analytical solution for the farmer's optimal land allocation decision rule at time t. Observe that μ_i's coefficients are some nonlinear function of λ_1, β, d_1, and δ_i's coefficients, which expresses the restriction imposed across the decision rule and the parameters of the stochastic processes for variables in $\mathbf{Z}_t$.

The certainty equivalence property of the model implies that under certainty the optimal decision rule (9) depends on the actual values (perfect foresight) of a_{1t+j} and R_{t+j} ($j = 0, 1, \dots$) and that under uncertainty it depends on their conditional expected values but not on other moments of the distributions of the random variables. Imposing the uncertainty on the farmer's problem is of interest here for two reasons. First, theoretically, past exogenous variables are correlated with current decisions only because of the information t[illegible] contain about their future realizations (as in [11]). Cha[illegible] distributions of the exogenous variables would alter the [illegible] through the cross-equations restrictions. Hence, predicti[illegible] spect to a permanent change in relative prices require[illegible] identification of the model's parameters, even though pr[illegible] enous (Lucas 1976). Second, under uncertainty equation ([illegible] standard econometric form of distributed lag models that are [illegible] mable using standard nonlinear methods. On the other hand, under certainty equation (9) contains all future exogenous variables, and it

---

[10] Hansen and Sargent's (1980) optimal projection for (9) given (8) and the information set Ω_{t-1}, is

$$\sum_{j=0}^{\infty} \lambda^j E_{t-1}(c_{t+j}) = \mathbf{v} \cdot \left\{\lambda^{-1}\delta^{-1}(\lambda)\left[I + \sum_{j=1}^{k-1}\left(\sum_{i=j+1}^{k} \lambda^{i-j}\delta_i\right)L^j\right] - \lambda^{-1}I\right\} Z_{t-1},$$

where $\mathbf{v} = [1, 1, 0, 0, \dots, 0]$ is a row vector with ones in the first two positions and zeros in the next $n - 2$ positions, and where $\mathbf{v} \cdot Z_t = c_t = (a_{t}, -R_t)$, $\lambda = \lambda_1\beta$, and I is an $n \times n$ identity matrix.

cannot provide an integrated aggregate error term for estimation (see Sec. III). Therefore, it is not a tractable estimable equation.

In order to analyze (11) let us consider the following case. The shocks to production (a_1's) and the price (R) are serially uncorrelated and are independent of variables that are in the information set; a_{1t} has zero mean and R_t has a positive mean. Equation (11) can be written as

$$A_{1t} = \lambda_1 A_{1t-1} + \frac{\lambda_1 \bar{A}}{d_1(1 - \lambda_1)} \cdot (\text{mean of } R), \tag{12a}$$

and the mean of A_{1t}, A_1^*, is

$$A_1^* = \frac{\gamma}{1 - \lambda_1} + \frac{\lambda_1 \bar{A}}{d_1(1 - \lambda_1\beta)(1 - \lambda_1)} \cdot (\textit{mean of } R). \tag{12b}$$

For the relevant domain of d_1, we obtain $\partial A_{1t}/\partial d_1 < 0$ and $\partial A_1^*/\partial d_1 < 0$. Thus, increasing the rate of land deterioration decreases the area allocated to crop 1. Equations (12a) and (12b) show that farmers would not interpret price fluctuations and shocks to production that are serially uncorrelated as signaling a permanent alteration in the incentives confronting them. Now consider the experiment of a once-and-for-all increase in the mean of the relative price, R. Using equations (12a) and (12b) the immediate response for A_{1t} is a decrease below the (lower) new level of A_1^* and, by oscillating fluctuations, convergence toward the new mean of A_{1t}. Hence, the short-run effect is greater than the long-run, and the cobweb phenomenon is, in this model, *an optimal response to a shock in relative prices and* has nothing to do with rational price expectations. At the same time, the model predicts the same response to a shock as the VAR (fig. 1) using Egyptian data on cotton and wheat. Observe that the qualitative implication of this experiment holds in the general case as well and can illustrate the effects of deterministic tax (subsidy) policies on the movements of land allocations. It is straightforward to see that in the case of adjustment costs ($\lambda_1 > 0$ and $d_1 < 0$), the signs of both the immediate and the long-run effects of the experiment are retained but the magnitude of both increases. However, the short-run effect is *lower* than the long-run (see Nerlove 1958) and the convergence toward the mean is a downward smooth path rather than oscillations as in the case where $d_1 > 0$. In general, the structure of the stochastic process of the relative price has an important effect on the predicted movements of land allocations due to changes in prices or other variables that affect prices. This includes the magnitude of the difference between the immediate response (short run) and the average change (long run) in land allocations due to changes in prices.

What are the effects of fertilizer, labor, and pesticides on the land allocation decision rule? In general, if the production function of crop I is *separable* between land and any other inputs, the decision rule (11) stays the same. Theoretically, one can specify a production function that exhibits a complicated interaction between factors of production and that includes both static and dynamic elements. Hansen and Sargent (1981) discuss methods for solving these types of models.[11]

Estimating the underlying parameters of the model is crucial in understanding supply responses and the land allocation decision process. Equation (11) is almost a regression equation. If I do not observe some of the variables that are part of the farmer's information set, I can construct an error term for (11) such that it has the properties of a regression equation. The reduced form of this equation is observationally equivalent to the traditional supply response model of Nerlove (1958), since both models can be specified such that they have identical reduced-form equations. However, the model in this paper has a completely different interpretation of the observed pattern of serial and cross correlations between crop areas and crop prices.[12] The traditional supply response model does not investigate jointly the dynamics of the production process and the dynamics of the actual crop prices that farmers observe. The model here emphasizes the important role of the dynamic structure of the production technology, the information farmers have at the time inputs are committe production, and the way relative prices are moving over determination of farmers' response to changes in incen

III. Estimating the Dynamic Land Allocation Model

The reason for estimating the model is to determine whe namic rational expectations model is reliable for prediction and capa ble of interpreting the observed serial and cross correlations between crop areas and prices. In the discussion above I showed that the model can interpret oscillating fluctuations in land allocations, such as those that arise in a VAR (fig. 1) using Egyptian data, as a rational response to shocks in prices and productivity. As such, there is reason to believe that the restrictions imposed on the data by the model will not be rejected by a careful use of statistical methods.

Here I present results from a maximum likelihood estimator of a simple bivariate specification of the land allocation model using Egyptian annual data on cotton area, cotton lint price, wheat price, and wheat yield from 1913 to 1969. Cotton and wheat are the main crops in the Egyptian economy, and soil deterioration and insect accumulation in soils under continuous cotton production are the main reasons for crop rotation in Egyptian agriculture. During the relevant period both cotton and wheat were internationally traded in large quantities. Hence, it is reasonable to assume that their prices are determined in the world markets and are not Granger caused (Granger 1969) by Egyptian production. The average area and yield of cotton and wheat stayed almost the same over the entire period, but the annual quantities show large fluctuations over time. Given the importance of agriculture in the poor Egyptian economy, the characteristics of the main crops, and the fact that most recent studies on the dynamics of agricultural supply were of crops in less developed countries (see Askari and Cummings 1976), it is particularly interesting to test this model with the fairly good Egyptian data.[13]

---

[11] The App. illustrates how the interaction between inputs may affect the dynamic properties of the model.

[12] See Sargent (1976) for a similar result with respect to macroeconomic models. In Eckstein (1981*a*) I show how a simple Nerlovian model can fit perfectly well data generated from a specific case of eq. (11).

As in traditional agricultural supply response models, the two variables here are the cotton crop area (A_{1t}) and the relative price (R_t). The decision rule (11) and the vector stochastic process (8) are the joint equations that are estimated here. The unrestricted (reduced-form) system of two equations (A_{1t}, R_t) yields a VAR of a lag length that is determined by the stochastic process for R_t and a_{1t}. Therefore, an unrestricted VAR can serve as a statistically maintained alternative to any restrictions that the theory imposes on the data. Likelihood ratio tests of lag length for the complete unrestricted (symmetric lags) rejected two versus four lags (marginal confidence level = .92) but did not reject three versus four lags (marginal confidence level = .44).

Two types of identifying restrictions exist here: (i) zero restrictions, which amount to the assumption that A_{1t} does not Granger cause R_t, and (ii) cross-equation restrictions due to the rational expectations assumption, given that i is correct (see Sargent 1981).

---

[13] Almost identical data were used by Hansen and Nashashibi (1974, 1975) and are available also in Eckstein (1981*b*). The prices are ex-farm prices in Egypt and the areas and yield are national aggregates The sources of the data are Egypt Ministry of Finance, *Annuaire Statistique*, 1909 60, and National Bank of Egypt, *Economic Bulletin*, 1948–78 Detailed discussions of the Egyptian agricultural sector are available in Hansen and Marzouk (1965), Owen (1969), Hansen and Nashashibi (1975), and Eckstein (1981*b*) Note that cotton is the main crop in production, and both the lint and the seeds have been the main sources of export earnings for many years (since 1880). Wheat is second to cotton in production; its growing period overlaps cotton's, and it is a part of the crop rotation system that Egyptian farmers follow. Furthermore, wheat became an important imported commodity, and substitution between wheat and cotton in production has a direct effect on the trade balance

In order to preserve the simplicity of the model and the three-lag VAR supported by the tests, I assume that R_t and the shocks to productivity (a_{1t}) have the following autoregressive processes:[14]

$$\begin{cases} R_t = \alpha_0 + \alpha_1 R_{t-1} + \alpha_2 R_{t-2} + u_t^R \\ a_{1t} = \rho a_{1t-1} + u_t^a, \end{cases} \tag{13}$$

where it is assumed that $|\rho| < 1$ and the roots of $|1 - \alpha_1 z - \alpha_2 z^2| = 0$ are outside the unit circle.

Using the farmer's land allocation decision rule (11), and given that a_{1t} is not observed, I can write the VAR for A_{1t} and R_t as

$$\begin{bmatrix} A_{1t} \\ R_t \end{bmatrix} = \begin{bmatrix} \mu_0 \\ \alpha_0 \end{bmatrix} + \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \text{WAR} + \begin{bmatrix} \rho + \lambda_1 & \mu_1 \\ 0 & \alpha_1 \end{bmatrix} \begin{bmatrix} A_{1t-1} \\ R_{t-1} \end{bmatrix} \tag{14}$$

$$+ \begin{bmatrix} -\rho\lambda_1 & \mu_2 - \rho\mu_1 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} A_{1t-2} \\ R_{1t-2} \end{bmatrix} + \begin{bmatrix} 0 & -\rho\mu_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A_{1t-3} \\ R_{1t-3} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix},$$

where WAR represents dummies for the Second World War (1941–45),[15] μ_0 contains several deterministic (time-independent) elements from the decision rule (11) that cannot be identified separately ϵ $\mu_3 u_{t-1}^a$, and $\epsilon_{2t} = u_t^R$; μ_1, μ_2, and μ_3, defined in (15) restrictions across the equations in (14) and represent the of the rational expectations hypothesis:[16]

$$\mu_1 = \frac{\lambda_1}{d}\left(\frac{\alpha_1 + \alpha_2\lambda}{1 - \alpha_1\lambda - \alpha_2\lambda}\right)$$

$$\mu_2 = \frac{\lambda_1}{d}\left(\frac{\alpha_2}{1 - \alpha_1\lambda - \alpha_2\lambda^2}\right)$$

$$\mu_3 = \frac{\lambda_1}{d}\left(\frac{\rho}{1 - \rho\lambda}\right),$$

where $d = d_1/\bar{A}$. Here $\boldsymbol{\epsilon}_t = (\epsilon_{1t}, \epsilon_{2t})'$ is the vector of innovations that is assumed to have a bivariate normal distribution with $E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t') = V$.[17]

---

[14] R_t = wheat price × wheat yield × (1/cotton lint price)

[15] The Second World War caused exogenous limitations on trade and altered the level of wheat and cotton production To capture this effect a dummy was introduced to the empirical specifications

[16] See n. 10 for the method.

[17] Equation 13 is a particular specification for eq. (8) The lag order in the R_t process is supported by estimating univariate autoregressive processes. Strategies for estimat-

Hence, estimators of the vector of free parameters $\boldsymbol{\theta} = \{\lambda_1, d, \rho, \alpha_1, \alpha_2, \alpha_0, \mu_0, w_1, w_2\}$ is obtained by maximizing the likelihood function with respect to $\boldsymbol{\theta}$. Let $\mathbf{l}_t = (l_{1t}, l_{2t})'$ be the sample residual vector for a given value of the parameter vector $\boldsymbol{\theta}$. Then the concentrated log-likelihood function of the sample of observations on the residuals over $t = 1, \ldots, T$ is (see Sargent 1978)

$$L(\boldsymbol{\theta}) = -T[\log(2\Pi) + 1 - \log T] - \tfrac{1}{2}T \log \left| \sum_{t=1}^{T} \mathbf{l}_t(\boldsymbol{\theta})\mathbf{l}_t(\boldsymbol{\theta})' \right|. \qquad (16)$$

Equation (16) is maximized with respect to $\boldsymbol{\theta}$, where $\mathbf{l}_t(\boldsymbol{\theta})$ is defined by (14) and (15) for each observation.[18] Observe that (14) has 11 nonzero regressors while the vector $\boldsymbol{\theta}$ has only nine free parameters. Hence, there are two overidentifying restrictions that are imposed by theory as implied by (15). These restrictions as well as the a priori zero restrictions are tested using conventional likelihood ratio tests.

The estimated parameters (see table 1) of the model satisfy the restrictions that the model imposed on the farmer's problem; that is, $|\lambda_1| < 1$, $|\rho| < 1$, the roots of $|1 - \alpha_1 z - \alpha_2 z^2| = 0$ are outside the unit circle, and the sign of d is opposite to the sign of λ_1. The values of λ_1 and d are consistent with adjustment cost effects in production and are not compatible with the simple specification of the soil fertility deterioration in cotton production. In the Appendix I show that the interaction between land and fertilizer may affect the dynamics of the land allocations decision in such a way that if data on fertilizer are omitted λ_1 may be positive. Thus, the traditional omitted-variable argument may explain the signs of λ_1 and d_1, maintaining the hypothesis of soil fertility deterioration in cotton. Using the estimated parameters I can calculate the response of land allocations to a permanent or temporary change in prices. The long-run supply elasticity, that is, the percentage change in the mean of A_1 divided by the percentage change in the mean of R, is equal to $-.13$; and the immediate effect of an expected change in R_t (short-run elasticity) has a

---

ing this type of model are discussed in Sargent (1978), Hansen and Sargent (1980), and Wallis (1980) The time domain full-information maximum likelihood is the most conventional method for multivariate nonlinear models. It turns out that for our model this method is also computationally more efficient than frequency domain approximations of the likelihood function. In Eckstein (1981*b*) a four variate model has been estimated using frequency domain approximations to the likelihood function.

[18] The maximization was done using the DFP algorithm from the GQOPT package of Princeton University. The complicated nonlinear structure of the model implies no gain from writing the analytical first and second derivatives; hence, I used the derivative-free method I maintained a 10-digit accuracy level and checked for other local maxima in "the neighborhood." I do not report the asymptotic standard errors of our estimators since the Hessian, at the maximum, was not negative definite. The computer program had been tested using a Monte Carlo experiment of the same model that I estimated.

TABLE 1

ESTIMATED PARAMETERS OF THE LAND ALLOCATION MODEL

$\lambda_1 = .081$	$\mu_0 = 1,551.03$
$d = -.008$	α_0 3.79
$\alpha_1 = .524$	$w_1 = -719.13$
$\alpha_2 = .250$	$w_2 = .06$
$\rho = .081$	

NOTE.—The log likelihood = $L(\theta) = -506.088$. β = discount factor = .95, imposed a priori.

lower value ($d < 0$) that is equal to $-.11$.[19] As such, all the estimated parameters of the model have signs consistent with the assumptions of the model, and the magnitudes of the long-run and the short-run elasticities are within a reasonable range.

Let $L(\theta)$ be the value of the log likelihood of the model and L_u be the value of the log likelihood of an estimated unrestricted version of the VAR (14). Then, $-2[L(\theta) - L_u]$ is distributed $\chi^2(q)$, where q is the number of restrictions that are tested. Three likelihood ratio tests of the model were performed. First, the null hypothesis that the parameters λ_1, d, α_1, α_2, and ρ (not the constants and the dummies) are zero is tested against the model. (This test is similar to the standard F-test of the linear regression equation.) The likelihood value of the null hypothesis is -529.63, and it is rejected with a marginal significance level that is less than .001. Hence, the implications of the model on th
pattern of serial and cross correlations between prices and
more consistent with the data than the null hypothesis of no
cross correlations.

Second, I test the overidentifying cross-equations restric
to the rational expectations solution of the decision rule, u
the assumption that A_1 does not Granger cause R; that is
restrictions on the three-lag VAR are not relaxed. The es
VAR implied by the model is presented at the top of table 2. T
estimated VAR where the cross-equation restrictions are not imposed is reported in the middle of table 2: the likelihood value of this system is -505.489. The marginal confidence level of testing the theory against the less restricted model is less than .5 ($\chi^1[2] = 1.2$), and the model is not rejected. Table 2 indicates that the coefficients of the restricted and unrestricted (with zero restrictions) VARs have the same sign as the auto- and cross covariances. The land equation shows a higher negative effect in the second-order lags of A_1 and R, and the

---

[19] The average of R = 16.8. The average of A_1 = 1,530.0. The long-run elasticity is

$$\frac{\lambda_1}{d(1 - \lambda_1\beta)(1 - \lambda_1)} \cdot \frac{16.8}{1,530} = -.13$$

The immediate elasticity (short run) = $(\lambda_1/d)(16.8/1,530) = -.11$

TABLE 2

THE ESTIMATES OF THE VAR (14)

Model's VAR, $L(\theta) = -506.088$.

$$\begin{bmatrix} A_{1t} \\ R_t \end{bmatrix} = \begin{bmatrix} 1{,}551.0 \\ 3.8 \end{bmatrix} + \begin{bmatrix} -719.1 \\ .66 \end{bmatrix} \text{WAR} + \begin{bmatrix} 16 & 5.7 \\ 0 & .52 \end{bmatrix} \begin{bmatrix} A_{1t-1} \\ R_{t-1} \end{bmatrix} + \begin{bmatrix} -.006 & -2.1 \\ 0 & .25 \end{bmatrix} \begin{bmatrix} A_{1t-2} \\ R_{t-2} \end{bmatrix} + \begin{bmatrix} 0 & -.21 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A_{1t-3} \\ R_{t-3} \end{bmatrix}$$

Unrestricted VAR (with zero restrictions), $L_{u1} = -505.489$:

$$\begin{bmatrix} A_{1t} \\ R_t \end{bmatrix} = \begin{bmatrix} 1{,}563.5 \\ 3.9 \end{bmatrix} + \begin{bmatrix} -724.5 \\ .81 \end{bmatrix} \text{WAR} + \begin{bmatrix} 19 & -3.8 \\ 0 & .56 \end{bmatrix} \begin{bmatrix} A_{1t-1} \\ R_{t-1} \end{bmatrix} + \begin{bmatrix} -.06 & -7.2 \\ 0 & .21 \end{bmatrix} \begin{bmatrix} A_{1t-2} \\ R_{t-2} \end{bmatrix} + \begin{bmatrix} 0 & -4.4 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} A_{1t-3} \\ R_{t-3} \end{bmatrix}$$

Unrestricted VAR (without zero restrictions), $L_{u2} = -495.9$

$$\begin{bmatrix} A_{1t} \\ R_t \end{bmatrix} = \begin{bmatrix} 1{,}197.6 \\ -14.2 \end{bmatrix} + \begin{bmatrix} -661.4 \\ 4.6 \end{bmatrix} \text{WAR} + \begin{bmatrix} 27 & -7.6 \\ .005 & .36 \end{bmatrix} \begin{bmatrix} A_{1t-1} \\ R_{t-1} \end{bmatrix} + \begin{bmatrix} -.11 & -8.7 \\ .002 & .11 \end{bmatrix} \begin{bmatrix} A_{1t-2} \\ R_{t-2} \end{bmatrix} + \begin{bmatrix} .21 & 9.7 \\ .004 & .36 \end{bmatrix} \begin{bmatrix} A_{1t-3} \\ A_{t-3} \end{bmatrix}$$

patterns of the moving average coefficients of the two VARs are almost identical. This test can be viewed as a support for the rational expectations solution versus any econometric model that ignores the cross-equation restrictions but that is observationally equivalent (Sargent 1976) to the model presented here. The Nerlovian supply response model is an example of an econometric model that for some specifications has the same VAR as for the model presented here.[20]

Third, the likelihood ratio test of the model versus a three-lag unrestricted VAR (the third system in table 2) rejects the null hypothesis with a marginal confidence level of .995 ($\chi^2[7] = 20.3$). The signs of the serial and cross correlations of the area equation are the same as in the model, but the magnitude is significantly different. The main change is in the R_t equation where the zero restrictions are relaxed. The serial correlations of R have the same sign and only marginally different values compared with the VAR of the model. Hence, the model is rejected because of the zero restrictions on the price equation and not because of the rational expectations cross-equations restrictions. The result indicates that the model probably fails because of the way I assume prices are determined in the Egyptian economy rather than because of the particular linear-quadratic approximation of the dynamic technology or the implied rational expectation restrictions. Given a priori acceptable point estimates and the positive outcome of the second test, I can characterize the empirical results as providing some support for the specific mod

IV. Conclusion

This work is best viewed as an attempt to construct a dynam tic theory of agricultural supply that can be used to interpre data on land allocations, crop yields, and prices. By introdu explicit approximation to a well-known characteristic of the crop production process (depletion of soil productivity), I demonstrate how the dynamic properties of land allocation and their interaction with crop prices depend on the production technology. The model can interpret the dynamics of land allocations as a result of different technologies: the depletion effect in land productivity; costs of adjusting crop areas; or omitted inputs that interact with land (e.g., fertilizer). In such a model the supply response elasticities are functions of the technology and the parameters of the price processes. It turns out that the restrictions imposed by the model received some support from the data. The estimated structural parameters are consistent with the model and conform to a cost-of-adjustment framework, and the cross-equations restrictions are not rejected when the zero restrictions on the price equation are maintained. The methodology adopted here requires that future changes in the model demand careful consideration of aspects of the agricultural decision process that have been ignored previously. As such, there is hope that future research will lead us to an improved understanding of the dynamics of agricultural supply.

---

[20] See Askari and Cummings (1976) for reported results from the Nerlovian model

Appendix

Here I consider a simple example with fertilizer. Let F_{1t} be the fertilizer that is allocated to crop 1 at time t, and let the production function for crop 1 be

$$X_{1t} = \left(f_1 + a_{1t} - \frac{g_1}{2}A_{1t}\right)A_{1t} + d_1\left[w_1F_{1t} - \left(\frac{A_{1t-1}}{\bar{A}} - \frac{A_{1t}}{\bar{A}}\right)\right]A_{1t} - \frac{w_2}{2}F_{1t}^2 \tag{A1}$$

where w_1 and w_2 are positive scalars. Substituting (A1) rather than (5) into (3) and subtracting the cost of fertilizer from the farmer's problem, I can find the first-order necessary conditions of the farmer's problem with respect to F_{1t} and A_{1t}. Solving for F_{1t} in terms of A_{1t} and the current price of fertilizer and substituting the result into the first-order condition with respect to A_{1t}, I find that the land allocation decision rule has exactly the same form as the solution for the original problem (7). Here the price of fertilizer is an additional element in the optimal decision rule and in the uncontrolled vector stochastic process of $\mathbf{Z}_t$. However, there exists an important difference between the two solutions. If $d_1^2w_1^2/w_2 > g_1 + 2(d_1/\bar{A})$, the coefficient of A_{1t} in the Euler equation is negative and it is possible to have a real solution with $0 < \lambda_1 < 1$. Hence the serial correlation in land allocation is positive, such as in the adjustment costs case. The intuitive interpretation of this result is that if the production of crop 1 is very responsive to fertilizer applications (large w_1 and small w_2), the rotation element in land allocation may completely disappear.

References

Askari, Hossein, and Cummings, John T. *Agricultural Supply Response: A Survey of the Econometric Evidence* New York: Praeger, 1976.

Behrman, Jere R. *Supply Response in Underdeveloped Agriculture: A Case Study of Four Major Annual Crops in Thailand, 1937–1963* Amsterdam: North-Holland, 1968.

Eckstein, Zvi. "Rational Expectations Modeling of Agricultural Supply." Economic Growth Center Discussion Paper no. 381. New Haven, Conn.: Yale Univ., August 1981. (*a*)

———. "Rational Expectations Modeling of Agricultural Supply: The Egyptian Case." Ph.D. dissertation, Univ. Minnesota, 1981. (*b*)

Gould, John P. "Adjustment Costs in the Theory of Investment of the Firm." *Rev. Econ. Studies* 35 (January 1968): 47–55.

Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37 (July 1969): 424–38.

Hansen, Bent, and Marzouk, Girgis A. *Development and Economic Policy in the UAR (Egypt)*. Amsterdam: North-Holland, 1965.

Hansen, Bent, and Nashashibi, Karim. "Protection and Competitiveness in Egyptian Agriculture and Industry." Working Paper no. 48. Cambridge, Mass.: Nat. Bur. Econ. Res., 1974.

———. *Foreign Trade Regimes and Economic Development. Egypt.* New York: Columbia Univ. Press (for Nat Bur. Econ. Res.), 1975

Hansen, Lars P., and Sargent, Thomas J. "Formulating and Estimating Dynamic Linear Rational Expectations Models." *J Econ. Dynamics and Control* 2 (February 1980): 7–46.

———. "Linear Rational Expectations Models for Dynamically Interrelated Variables." In *Rational Expectations and Econometric Practice*, edited by Robert E. Lucas, Jr. and Thomas J. Sargent. Minneapolis: Univ. Minnesota Press, 1981.

Lucas, Robert E., Jr "Adjustment Costs and the Theory of Supply." *J P E.* 75, no. 4, pt. 1 (August 1967): 321–34.

———. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, edited by Karl Brunner and Allan H Meltzer Carnegie-Rochester Conference Series on Public Policy Amsterdam North-Holland, 1976.

Marschak, Jacob. "Economic Measurements for Policy and Predictions" In *Studies in Econometric Method*, edited by William C. Hood and Tjalling C. Koopmans. Cowles Monograph no. 14. New York: Wiley, 1953

Muth, John F. "Rational Expectations and the Theory of Price Movements" *Econometrica* 29 (July 1961): 315–35.

Nerlove, Marc. "Estimates of the Elasticities of Supply of Selected Agricultural Commodities." *J. Farm Econ.* 38 (May 1956): 496–509

———. *The Dynamics of Supply Estimation of Farmers' Response to Price* Baltimore: Johns Hopkins Univ. Press, 1958.

———. "Lags in Economic Behavior" *Econometrica* 40 (March 1972) 221–51

———. "The Dynamics of Supply: Retrospect and Prospect" *American J Agricultural Econ* 61 (December 1979): 874–88

Nerlove, Marc; Grether, David M.; and Carvalho, José L. *Analysis of* *Time Series: A Synthesis.* New York Academic Press, 1979

Owen, Edward R. J. *Cotton and the Egyptian Economy, 1820–191* *Trade and Development.* Oxford: Clarendon, 1969.

Sargent, Thomas J "The Observational Equivalence of Natur natural Rate Theories of Macroeconomics." *J P E* 84 (June 197

———. "Estimation of Dynamic Labor Demand Schedules und Expectations." *J.P.E* 86 (December 1978). 1009–44.

———. "Causality, Exogeneity, and Natural Rate Models Reply . Nelson and B. T. McCallum." *J P.E* 87 (April 1979): 403–9 (*a*)

———. *Macroeconomic Theory* New York: Academic Press, 1979. (*b*)

———. "Interpreting Economic Time Series." *J P E* 89 (April 1981): 213–48.

Schultz, Theodore W., ed. *Distortions of Agricultural Incentives* Bloomington Indiana Univ. Press, 1978.

Sims, Christopher A. "Macroeconomics and Reality." *Econometrica* 48 (January 1980): 1–48.

Wallis, Kenneth F. "Econometric Implications of the Rational Expectations Hypothesis." *Econometrica* 48 (January 1980): 49–73.

The Taxation of Risky Assets

Jeremy I. Bulow
Stanford University

Lawrence H. Summers
Harvard University

This paper reconsiders the effects of taxation on risky assets, recognizing the importance of variations in asset prices. We show that earlier analyses that assumed that depreciation rates are constant and that the future price of capital goods is known with certainty are very misleading as guides to the effects of corporate taxes. We then examine the concept of economic depreciation in a risky environment and show that depreciation allowances, if set ex ante, should be adjusted to take account of future asset price risk. Some empirical calculations suggest that these adjustments are large and have important implications for the burdens of, and nonneutralities in, the corporation income tax.

I. Introduction

The enormous volatility in the prices of capital assets in the American economy has been widely noted. The standard deviation of real stock market returns appears to be about 20 percent per year. Over the last 15 years the price of used capital goods as measured in the stock market relative to consumption goods has varied by a factor of more than 2. Comparable volatility is observed in the pricing of used capital goods such as airplanes and office buildings where an active secondhand market exists. It seems clear that realistic positive or normative analysis of the effects of capital income taxation requires models in which there is substantial scope for variation in the price of capital

[*Journal of Political Economy*, 1984, vol. 92, no. 1]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0010$01.50

assets. Yet the literature on capital income taxes and risk taking has focused almost entirely on models where the return from capital goods is variable but their relative price is certain.

This paper reconsiders the effects of taxes on the return from risky assets, recognizing the importance of variations in the price of capital goods. The results suggest that the burden of the corporate income tax is much greater than that implied by analyses such as Feldstein (1969), Stiglitz (1969), and Gordon (1981), which suggest that the government's risk-sharing role almost fully compensates investors for the taxes they pay. The essential error in earlier analyses is in the treatment of depreciation. The observation that capital asset prices are far more volatile than earnings streams suggests that most of the risk associated with capital assets is in their rate of economic (though perhaps not physical) depreciation rather than in their contemporaneous marginal product. The tax laws' use of ex ante depreciation schedules rather than ex post depreciation means that depreciation or "capital" risk is not shared by the tax collector. Therefore, a tax like the U.S. corporate income tax absorbs a much greater part of the return than of the risk on capital investments.

The implications of capital risk for depreciation policy are then considered. We argue that traditional concepts of what constitutes economic depreciation are likely to lead to serious errors in realistic settings. In particular, the notion that pervades theoretical and empirical work on depreciation policy, that economic depreci
be measured by the expected decline in the price of capit
shown to be wrong. We derive a new measure of economic
that can be calculated from observable market data on ass
and prices and use it as a benchmark for evaluating c
policies.

The consequences of using this new measure of economic durability for a number of tax policy questions are analyzed. Using two alternative empirical methods based on financial market data, we show that prior calculations of appropriate economic depreciation represent serious underestimates because of their failure to consider risk. We also show that proper risk accounting suggests that previous studies of the problem of neutrality between assets of different durability are seriously incomplete. A final implication of the results is that true rates of economic depreciation have increased substantially over the last decade due to increasing uncertainty.

The next section of this paper examines the effects of capital income taxes on investment within standard models where the price of capital goods is known with certainty. A number of serious logical and empirical problems that arise in applying this analysis to actual corporate income taxes are then pointed out. The crucial distinction between accounting income, on which the corporate tax is levied, and real economic income, which includes capital gains and losses, is emphasized. The third section shows that these difficulties can be avoided by recognizing the distinction between income and capital risk. The observation that most risk is of the latter variety leads to a reconsideration of tax depreciation policy. In the fourth section our new economic durability measure is introduced and the concept of economic depreciation in a risky environment is analyzed. The implications of the analysis for the evaluation of tax policy are taken up in the fifth section. Empirical estimates based on financial market data, suggesting the importance of taking account of asset price risk in estimating rates of economic depreciation, are also presented. A sixth section of the paper summarizes the results and suggests a direction for subsequent theoretical and empirical research.

II. Taxes and Risk Taking When Depreciation Is Known with Certainty

This section briefly reviews previous results regarding the effects of taxation on investment in risky assets. The paradoxical conclusions obtained in much of the literature suggesting that taxes on risky assets may actually encourage investment in them are shown to depend on implausible assumptions. In particular, these conclusions depend on the government's presumed ability to bear risk costlessly and on the assumption that the relative prices of capital assets do not vary. The implausibility of these assumptions is illustrated by noting that they imply that the market value of the claim taken by the government through the corporate tax system is negligible.

We begin by reviewing a simple model of corporate investment in a mean variance setting similar to the one employed in earlier work on taxation and risk taking. The perfect markets assumptions necessary for the validity of the capital asset pricing model (CAPM) are presumed to hold.[1] Individuals are compensated only for the systematic risk they bear. Diversifiable risk is not compensated so that private and social risk are equal. This implies that there is no gain from risk spreading through government taxation of risky assets. The private capital market is able to accomplish all feasible risk sharing.[2]

Now assume that the cash flow yielded by an asset fluctuates, but the price of capital assets is constant, and that assets depreciate at a constant rate that is known with certainty. We also assume that the taxed sector is sufficiently small that it is a price taker in determining the after-tax riskless rate and the slope of the capital market line (i.e., the risk-return trade-off) available to investors.[3] This assumption greatly simplifies the analysis without altering qualitative results. In the context of the CAPM in the absence of taxes, the corporate sector will invest up to the point where

---

[1] These include the absence of transactions costs, limitations on short selling, homogeneous expectations, the existence of a safe asset, and competitive behavior.

[2] If these assumptions are not satisfied, the government can increase welfare by serving as a financial intermediary. However, if there are economic reasons for the nonexistence of markets, such as moral hazard problems, there is no presumption that tax policy can increase welfare.

$$f'(K)^e - \delta = r + \alpha, \tag{1}$$

where $f'(K)^e$ is the expected marginal product of capital, assumed to be a declining function of K, δ is the rate of depreciation, r is the real interest rate, and α is the risk premium on corporate sector investments. In the context of the CAPM, $\alpha = \text{cov}\,\{[f'(K), r_M]/\text{var}\,(r_M)\} \cdot \bar{r}_M$ where r_M is the excess rate of return on the market portfolio and $\bar{r}_M$ is its expected value. Note that because of the assumption of a constant price of capital, *no* term reflecting capital gains or losses appears in (1). This assumption is relaxed below.

Now introduce a tax at rate τ on the risk premium earned on investments. That is, the government taxes all corporate income but allows deductions of $(\delta + r)K$ to cover depreciation and the riskless return on capital. Because the government is bearing 100τ percent of the risk on corporate investment as well as 100τ percent of the risk premium, investors face the same capital market line (risk-return trade-off) as before the tax. Capital market equilibrium requires

$$r + \delta + (1 - \tau)[f'(K)^e - r - \delta] = r + \delta + \alpha(1 -$$

$$f'(K)^e - \delta = r + \alpha,$$

just as with no tax.

In models such as those of Tobin (1958) and Stiglitz
assumed that any risk borne by the government is not pass
the private sector. In those models, investors will hold *more* risky
in the presence of a tax on the risk premium. An approach we favor, adopted by Gordon (1981) and Mintz (1981), assumes that this risk must be passed back to taxpayers. Assuming that the risk does stay in the economy and is therefore passed back to investors, it is easy to show, as Gordon has, that the level of risky investment is unaffected by a tax on the risk premium.

Hence, the risk premium tax yields expected revenues of $\alpha(1 - \tau)K$ but does not change the amount of risky investment. The reason is that while the government takes 100 percent of the risk premium, it also takes 100 percent of the risk. The slope of the capital market line facing investors is unchanged. The government is not distorting investment because it is extracting a claim of zero market value—its tax claim is the equivalent of its acquiring stock market futures contracts.

---

[3] For a defense of this proposition, see Summers (1981), who notes that the U.S. corporate sector represents less than one-tenth of world free market wealth

Now consider a pure income tax, where only depreciation deductions are allowed. With the income tax, if the posttax required rate of return remains the same, capital market equilibrium requires that

$$(1 - \tau)[f'(K)^{r} - \delta] = r + \alpha(1 - \tau) \tag{3}$$

or, equivalently,

$$f'(K)^{r} - \delta = \frac{r}{1 - \tau} + \alpha. \tag{4}$$

Given the very low level of real interest rates available in the American economy, this expression seems to imply that the corporate income tax has only a negligible impact on corporate investment. Assuming a 2 percent real interest rate and a Cobb-Douglas production function, this expression implies that the corporate income tax reduces the size of the capital stock by only about 13 percent.[1] This low estimate is obtained without considering the effects of accelerated depreciation, the investment tax credit, the tax advantages of debt finance, and any special ability of the government to spread risk via the tax. Despite the small increase in $f'(K)^{r}$ caused by the tax, the expected revenue yield is $\tau K[(r + \alpha)/(1 - \tau)]$, which may be quite large. Plausible American magnitudes suggest that expected annual tax revenues would represent about 5 percent of the market value of the capital stock.

However, though Gordon (1981) implies otherwise, there is no free lunch here. Because the government is taking a claim of zero market value, it can only finance programs whose budgets have a certainty equivalent of zero. If the government wishes to spend the expected revenues of this tax, it must also impose another tax, with zero expected revenue but a high market value (i.e., a countercyclical tax). The imposition of this second tax would be burdensome on the corporate sector and therefore decrease the level of investment in corporate risky assets.

Introducing inflation and personal taxes would not have any substantial effect on the preceding conclusions. The corporate tax in the foregoing analysis can be treated as an integrated tax including the effects of dividends and capital gains taxes. Introducing inflation would alter the conclusions slightly because of the taxation of nominal as well as real capital gains and interest. However, Gordon's (1981) analysis, which includes both inflation and personal taxes, reaches conclusions similar to those obtained here. The introduction of risk into the models that are normally used in public finance appears to have the dramatic implication that a corporate income tax has only very small allocative effects. In the next section we show that this is a consequence of the failure of standard models to account for fluctuations in the relative price of capital goods.

---

[1] This calculation assumes that the pretax gross marginal product of capital is .20 and the aggregate production function of the corporate sector is Cobb-Douglas with a capital share of .25. It is also assumed that labor is supplied inelastically to the corporate sector.

III. Fluctuations in the Price of Capital Goods

The foregoing discussion implies that a corporate income tax on economic income raises a substantial expected revenue without greatly increasing the required rate of return on (or decreasing the amount of) corporate investment. In this section we examine the applicability of analyses of the type reviewed in the previous section to the U.S. corporate income tax. We conclude that they are not applicable because the vast majority of the risk borne by corporate investors involves capital gains and losses as the relative price of corporate capital goods changes. The corporate income tax is levied on accounting income, which excludes these capital gains and losses. Hence it does not share in this type of risk.

This point can be clarified by distinguishing between an asset's "income risk" and "capital risk." The total return on an as
current return, f_K, plus its change in value, $\dot{p}_K$. Income ri
uncertainty in f_K, capital risk to uncertainty in $\dot{p}_K$. The as
virtually any corporate stock exhibit primarily capital risk.
ance in stock price is generally many times the variance i

The corporate tax shares in income risk because varia
rent economic income provide roughly equivalent variatio
able income. This risk sharing largely compensates investors for decrease in their expected return caused by the tax. Capital risk is another matter. The only allowance for changes in the value of capital goods is through certain *predetermined* depreciation deductions. While an asset that eventually becomes worthless will have the right amount of total depreciation deductions over its life, we will show that the mistiming of realization of income under the tax system causes a substantial bias against risky investment.

The implausibility of the standard framework's interpretation of the corporate income tax is easily exhibited. If corporate capital requires a significant risk premium, it follows that there must be a sizable risk that the return from holding corporate capital is less than the risk-free rate. Otherwise, this asset would dominate the safe asset.

The earnings-price ratio measures the rate of return investors would receive if the relative price of capital goods remained constant as assumed in previous analysis of the effects of taxation on risk taking. The lowest value of the earnings-price ratio observed since 1948 was 4.62 percent, far in excess of any estimate of the riskless rate.[5] An alternative way to view the problem is in terms of the after-tax net marginal product of capital. Holland and Myers (1979) report that the lowest value of this statistic was 3.6 percent in 1974. Many studies have estimated the pretax marginal product of capital and have found that it consistently lies above 8 percent. These figures imply that if the relative price of capital goods were fixed, corporate investments would dominate acquisitions of the safe asset. In order to make the same point in yet another way, note that the real interest rate has averaged close to zero. Yet, corporate tax collections have always been significantly positive, implying that the tax base has always been significant and positive. It seems clear that any effort to model taxes taking into account risks borne by investors cannot tell a consistent *story if the fiction that the relative price of capital remains constant is* maintained.[6]

The importance of changes in the relative price of capital goods may be seen directly in a number of ways. At the aggregate level, the q ratio of the market value of the corporate capital stock to an estimate of its "replacement cost" has varied between 0.56 and 1.24 over the past 20 years. This measure involves the hard-to-measure replacement cost of capital. Further evidence comes from examining the variance in stock market returns. Ibbotson and Sinquefield (1979) report that the standard deviation of annual real stock market returns is about 20 percent. Note that if the relative price of capital were constant, annual stock market returns would equal earnings-price ratios. The standard deviation of the earnings-price ratio is less than 3 percent, and the standard deviation of the marginal product of capital is less than 2 percent. This implies that variations in earnings account for between 1 percent and 17 percent of the variation in stock market returns.[7] It seems reasonable to conclude that most of the risk borne by corporate investors is capital risk, involving changes in the

---

[5] Similar conclusions can be obtained using the inflation-adjusted earnings-price ratios described in Summers (1983)

[6] There is a remote possibility of a "peso" problem, where the market risk premium reflects a low-probability disaster that has not yet taken place. Such a disaster would surely involve a change in the relative price of capital goods, as discussed below.

[7] The range reflects ambiguity in the allocation of the covariance term and the choice of concept Note that since much of the variation in earnings-price ratios and marginal products of capital are forecastable, these figures overstate the extent of earnings risk.

TABLE 1

STANDARD DEVIATIONS OF DEPRECIATION RATES

ASSET	AGE (Years)							
	1	2	3	4	5	6	7	8
Cars:								
Pinto		8.8	10.7	6.3	5.2			
Malibu		3.9	5.3	10.2	12.7			
Impala		3.1	3.7	6.9	14.1			
Trucks								
Ford F600	6.9	6.5	6.5	5.9	9.6	6.0	6.2	5.1
Ford C8000	1.2	3.2	5.6	5.6	5.9	7.2	11.3	9
International Harvester 1600	6.7	5.8	6.4	6.5	7.6	7.8	14.5	13.3
Chevrolet CE61003	7.8	5.2	6.0	5.8	6.8	6.6	10.5	11.5
Dodge D600	6.2	5.6	6.4	4.8	5.7	6.6	7.3	13.1

NOTE.—Data on cars were kindly provided by James Kahn. Data on trucks were provided by Dean Amel. All numbers in the table are percentages at annual rates. Data on cars were for 1972–77 models. Data on trucks were for 1971–78 models, with the prices of the various models tracked through 1980.

price of their asset, rather than income risk, involving changes in the current return on assets.

The enormous volatility in the relative price of capital shows up clearly in the markets for used capital goods. Table 1 presents some information on the variability in depreciation rates, as inferred from used car and truck prices. The data are drawn from publis to used-asset prices and so undoubtedly understate the actual transactions prices. Since they reflect nationwide av also understate the extent of uncertainty about the rate o the price of any individual capital good. Nonetheless, the very significant year-to-year variations in the rate of real p For example, the annual rate of real price decline for 2-ye F600 trucks varied between 25.8 percent in 1971 and 7.8 1976. Overall, for most models and vintages of cars and standard deviation of annual rates of real price declin percent. It is reasonable to expect that because of their sho and easy reproducibility, used car and truck prices should less volatile than those of other capital assets.

While data are not available on the rents earned by owners of cars and trucks, it appears that most of the risk borne by holders of these assets is capital risk involving changes in the price rather than income risk involving changes in the current return. These data corroborate the inference drawn from aggregate data that most of the risk borne by corporate investors involves changes in the relative price of corporate capital rather than movements in the marginal productivity of capital.

Economic Depreciation

As we showed in the previous section, most of the risk involved in holding an asset is "capital risk"—that the market value of the asset changes over time. The prior analysis of capital markets has assumed no capital risk—that the entire risk premium was for bearing "income risk"—risk in the current period's profits. In this section, we will consider the other extreme, which is a much closer approximation to the truth—that all risk is capital risk. We show that if in each period taxpayers are permitted deductions (which may be negative) equal to the change in the value of their assets, the analysis presented in the previous section continues to be applicable. However, the U.S. corporate income tax and that of all countries does not employ an ex post depreciation scheme—rather it employs a system of ex ante depreciation under which an investor's depreciation allowances are determined on the day the asset is acquired. Setting ex ante depreciation deductions equal to expected depreciation is not equivalent to permitting ex post economic depreciation. Calculations such as those presented in Jorgenson and Sullivan (1981) that ignore the distinction between ex ante and ex post depreciation schemes are likely to significantly understate the burden of the corporate tax.

We begin by considering ex post economic depreciation in a setting with capital risk. Let δK = the actual decrease in the value of capital assets in a given period. For reasons of exposition, we start by assuming $\delta^e K$, the expected decrease, is equal to zero. With economic depreciation, even though there is capital risk, the government is still taking on 100 percent of the risk in corporate investment. Therefore, the condition necessary for stockholders to invest in the corporate sector is

$$(1 - \tau)[f'(K)^e - \delta^e] = r + \alpha(1 - \tau) \tag{5}$$

or, given the assumption that $\delta^e = 0$,

$$f'(K)^e = \frac{r}{1 - \tau} + \alpha. \tag{6}$$

This required rate of return is the same as in (4) where no capital risk was assumed.

Current Depreciation Law

Actual U.S. depreciation, however, is not economic depreciation. We have ex ante depreciation, following a fixed schedule. What if, for the above asset where $\delta^e = 0$, we allowed ex ante depreciation equal to expected economic depreciation of zero? Recalling that capital risk is the risk in $\dot{p}_K$, we note that this risk is not reduced by a proportional tax on income that does not symmetrically tax capital gains and losses on asset holdings. Therefore, investors require an after-tax rate of return of $r + \alpha$, and the requirement for capital market equilibrium is

$$(1 - \tau)f'(K)^e - \tau\delta^e = r + \alpha, \tag{7}$$

but with $\delta^e = 0$ we have

$$f'(K)^e = \frac{r + \alpha}{1 - \tau}. \tag{8}$$

It is immediately apparent from a comparison of (6) and (8) that the required return on capital investments is greater in the latter case, because the government does not share in the risk that firms' capital will depreciate rapidly.

Notice that in the case of an asset with positive tax depreciation, the only distortion from using an ex ante depreciation scheme arises from timing considerations. Ultimately the purchaser of a \$1.00 asset gets \$1.00 in depreciation deductions. Not adjusting for risk is similar to any other depreciation problem; it creates a problem of timing.

The next section shows that it is possible to determine what ex ante depreciation scheme would be equivalent to allowing ex post economic depreciation. Only in the very special case of an asset with no capital risk is it appropriate to allow ex ante depreciation at the expected rates of depreciation on an asset.

IV. Ex Ante Economic Depreciation

Making Ex Ante Depreciation as Favorable as Ex Post Depreciation

We can now show what depreciation rate ex ante would be neces- to reduce the effective tax rate on a risky asset to τ. We first show that when, as above, expected depreciation is zero, providing ex ante exponential depreciation at rate α makes the investor indifferent to economic depreciation (and makes the effective tax rate τ). We then use another approach to show that if expected exponential depreciation is at rate δ, then the firm is as well off with uncertain economic depreciation deductions expected to be δ and certain exponential depreciation at rate $\alpha + \delta$. These results may seem paradoxical as they imply that the certainty equivalent of the stream of depreciation tax shields is *less* than their expected value. The paradox is resolved by noting that the tax shield due to economic depreciation is a "negative beta" asset. When the market does well, economic depreciation is low, and, conversely, when the market does poorly, depreciation is substantial.

With zero expected depreciation we can confirm that ex ante depreciation at rate α is the equivalent of economic depreciation. Note that with tax shields equal to $\alpha\tau$ times the amount of investment, the equation for capital market equilibrium is

$$\begin{gathered}(1-\tau)f'(K)^e + \alpha\tau = r + \alpha \\ f'(K)^e = \frac{r}{1-\tau} + \alpha.\end{gathered} \tag{9}$$

For the case where depreciation is expected at rate δ, the fiction of the firm considering the sales of its assets in the futures market is instructive. Consider an asset with an initial value of one, expected depreciation of δ, and a required risk premium of α. Then the firm could lock in a sales price for one period from now of $1 - \delta - \alpha$. Equivalently, it is equally happy with certain depreciation of $\delta + \alpha$ or uncertain economic depreciation with an expected value of δ.

Now consider what price could be received by agreeing to sell the asset two periods from now. If the risk premium and expected depreciation rate are constant, then the firm would be able to lock in a sales price two periods from now of $(1 - \delta - \alpha)^2$. Equivalently, the firm is equally happy taking economic depreciation in the second period or taking certain depreciation of $(1 - \delta - \alpha) - (1 - \delta - \alpha)^2 = (\alpha + \delta)(1 - \delta - \alpha)$. The analysis could be continued to show that the certainty equivalent of the risky depreciation deductions in period t is $(\alpha + \delta)(1 - \delta - \alpha)^{t-1}$.

To show that such certain depreciation is ex ante as favorable as ex post economic depreciation, we verify that the required rate of return with the proposed stream of certain deductions equals the required rate of return with ex post economic depreciation. In both cases we verify that for a break-even investment $f' = [r/(1 - \tau)] + \delta + \alpha$ as in (4).

For the proposed certain deductions, the net present value of an extra dollar invested in the asset, including risky cash flows and riskless depreciation allowances, can be written as

$$\sum_{t=1}^{\infty} \frac{f'(K)(1-\tau)(1-\delta)^{t-1}}{(1+\alpha+r)^t} + \sum_{t=1}^{\infty} \frac{\tau(\alpha+\delta)(1-\alpha-\delta)^{t-1}}{(1+r)^t} - 1. \tag{10}$$

Equation (10) verifies that if the marginal product of capital, $f'(K)$, equals $\alpha + \delta + [r/(1 + \tau)]$ the expected after-tax gross rents discounted at the risky rate of $\alpha + r$ plus the certain depreciation tax shields discounted at the riskless rate make the net present value (NPV) of the investment equal to zero.

With ex post economic depreciation, we have

$$\mathrm{NPV} = \sum_{t=1}^{\infty} \frac{f'(K)(1 - \delta)^{t-1}(1 - \tau) + \delta\tau(1 - \delta)^{t-1}}{[1 + \alpha(1 - \tau) + r]^t} - 1. \qquad (11)$$

In (11) the numerator is the firm's expected after-tax income in each period while the denominator is the rate appropriate to discounting these flows, given that the government has taken 100 percent of the risk. As in (10) the required rate of return, $f'(K)$, pretax, to make the NPV equal to zero, is $\alpha + \delta + [r/(1 - \tau)]$—just as with the previously described certain deductions.

Ex Ante Depreciation Rates as a Measure of Economic Durability

Numerous commentators on depreciation policy (e.g., Hulten and Wykoff 1981) suggest the expected decline in the market value of an asset as a proxy for economic depreciation. Our work indicates that a better measure would be the expected depreciation rate plus the portion of the risk premium attributable to capital risk. (As indicated earlier, this is most of the risk premium.)

Incorporating the capital risk premium in the definition of ex ante depreciation leads to a more satisfactory measure of "economic durability." For example, other things held equal, a greater proportion of the purchase price of a very risky asset will be for near-term cash
than for a less risky asset. Therefore, one would natura
think of the risky asset as economically less durable.

Similarly, a change in the riskless interest rate affects th
risky asset as though the risk premium were part of the d
rate: a one percentage point increase in r decreases the pr
of an exponentially depreciating constant risk premium
$100/(r + \delta + \alpha)$ percent—that is, the "duration" of the asset is
$+ \alpha)$.

This analysis has important implications for the issue of tax neutrality between assets of different types. Our conclusion that ex ante depreciation rates should be equal to $\delta' + \alpha$ suggests that current tax rules are likely to discriminate against risky investments because they are not allowed higher depreciation rates. Our analysis also bears on the question of distortion between assets of different durability. Longer-lived assets are likely to be subject to greater asset price fluctuations than are shorter-lived assets. As a consequence, they may be discriminated against by the tax system. These issues will be explored in more detail in a subsequent paper.

V. Implications for Tax Policy Analysis

The foregoing analysis suggests that any realistic description of depreciation must recognize its stochastic character. In this section, we show how the concepts developed above can be used to produce empirical estimates of what ex ante tax depreciation schedule is required to correspond to ex post depreciation. Some estimates based on market data of the overall rate of depreciation of the capital stock are then presented. These are compared with standard BEA estimates. Finally, the implications of our results for the measurement of effective tax rates, and for analyses of the effects of taxation of the choice between assets with differing durability, are then considered.

We continue to rely on the approximation that all risk is capital risk. In this case the previous discussion demonstrated that it is appropriate to add the asset's risk premium to its expected rate of physical depreciation in order to determine the appropriate rate of ex ante depreciation. This suggests one empirical method of deriving estimates of appropriate depreciation schedules. If data on a time series of used-asset prices can be obtained, and if the assumptions of the capital asset pricing model are accepted, the economic depreciation rate for an asset can be estimated as

$$\delta_i^* = \delta_i^e + \beta_i(R_m - R_f), \tag{12}$$

where δ_i^* is the rate of ex ante depreciation, which is the certainty equivalent of ex post depreciation, δ_i^e is the expected rate of depreciation, β_i is the asset's beta, R_m is the return on the market, and R_f is the riskless rate.

Unfortunately, we are unaware of any data set extensive enough to permit estimation of β for any type of used assets. Therefore, it is difficult to use (12) as a basis for deriving estimates of economic depreciation for particular types of capital assets. However, it is possible to use (12) to make an approximate estimate of the depreciation rate on the total capital stock. Estimates of δ^e prepared by the BEA for the National Income and Product Accounts imply an average depreciation rate of 10.5 percent for the nonfinancial corporate equipment and structures in 1979. Hulten and Wykoff (1981) obtained lower estimates using data on used-asset prices. If it is assumed that the risk characteristics of corporate capital are like those of unlevered equity, the second term in (12) is equal to about 6 percent. Ibbotson and Sinquefield (1979) report that the risk premium on the stock market, $R_m - R_f$, has averaged about 9 percent over the last 50 years. The 6 percent figure is obtained by assuming a debt-to-market value of equity ratio of one-half as implied by statistics reported in Gordon and Malkiel (1981).

This calculation illustrates that taking account of risk has important implications. These estimates suggest that it raises the appropriate average ex ante rate of depreciation from 10 to 16 percent. Stated differently, if double declining balance depreciation is assumed, our risk adjustment reduces the appropriate average tax lives from 20 to 13 years.

There is an alternative way of calculating the appropriate ex ante depreciation rate on assets. Equation (1) implies that in the absence of taxation investment will take place up to the point where

$$f' - \delta^r = \alpha + r. \tag{13}$$

This suggests that the appropriate rate of ex ante depreciation on an asset $(\delta^r + \alpha)$ can be estimated as

$$f' - r = \delta^r + \alpha. \tag{14}$$

That is, the appropriate rate of ex ante depreciation for an asset is given by the difference between its rental price ratio and the risk-free rate. The analysis is more complex in the presence of taxation since part of the value of a capital asset represents the present value of the depreciation tax shields that it carries. We illustrate this below when a generalization of (14) is used to calculate the depreciation rate on the aggregate capital stock. Data on rental price ratios for different capital assets are not readily available. Gordon (1979) estimates rental price ratios of close to 25 percent implying depreciation rates of
proximately 22 percent on airplanes. This compares to the
of 7.5 percent and an estimate of 18.3 percent by Hulten a
(1981). A common rule of thumb in real estate is that prop
for 100 months' rent. This implies according to (14) a de
rate of 6 percent assuming that cash expenses includin
taxes plus the riskless interest rate add to 6 percent. Th
Income and Product Accounts use a much smaller depreciat
These examples provide further evidence that current measures
economic depreciation do not provide a good guide to appropriate tax policy, if emulating the effects of ex post depreciation is the desideratum.

For the economy as a whole, we can make some calculations about the adequacy of depreciation deductions. We can write the market value of the corporate sector as

$$MV_t = R_t + B_t, \tag{15}$$

where MV_t = market value of the corporate sector at time t, R_t = present value of after-tax cash flows if no future depreciation were allowed, B_t = present value of depreciation tax shields (equals tax rate times present value of depreciation deductions). We also have

$$R_t(1 + r + \alpha) = C_t + E(R_{t+1}) \tag{16}$$

$$E(R_{t+1}) \equiv R_t(1 - \delta^e), \tag{17}$$

where C_t = after-tax cash flows produced in period t, less the tax savings due to depreciation in the period. That is, (16) says that, including the cash payout C_t and the expected terminal value of the capital $E(R_{t+1})$, the investor must have an expected return of $r + \alpha$. Equation (17) simply defines the expected depreciation rate as δ^e.

Combining (15), (16), and (17) yields

$$\alpha + \delta^e = \frac{C_t}{MV_t - B_t} - r, \tag{18}$$

where $\alpha + \delta^e$ is our measure of certainty equivalent economic depreciation. Formula (18) allows us to use aggregate data to estimate what we call ex ante depreciation.

Table 2 shows the ex post, ex ante, and National Income and Product Accounts depreciation rates for the years 1950–79. The ex post depreciation rate is meant to measure the percentage decrease in the real market value of corporate physical assets that were held at the beginning of the year. The ex post rate was calculated as gross investment in physical assets by nonfinancial corporations, taken from the national balance sheets, less the increase in the real market value of physical corporate assets. For any given year, the market value of physical corporate assets was taken by adding the market value of nonfinancial corporate equity plus short-term and long-term debt and subtracting financial assets. The market value of short-term debt and the value of financial assets come from the national balance sheets; the market value of long-term debt is from Bulow and Shoven (1982), who took the national balance sheet numbers and multiplied them by the ratio of market to book value of New York Stock Exchange bonds. To calculate the increase in the real market value of physical assets in year t, the market value at the end of year t was reduced by the market value at the end of $t - 1$, times one plus the inflation rate for year t as measured by the GNP deflator. The ex post depreciation rate, reported in the table, was calculated as the depreciation number derived above divided by the market value of physical assets at the end of the prior year, times one plus the inflation rate.

Ex ante physical depreciation was calculated using (18). It is the after-tax cash flow less the value of depreciation tax shields for a given year, C_t, divided by the market value of physical assets at the end of the prior year, MV_t (calculated above), less the present value of all

TABLE 2

COMPARATIVE DEPRECIATION RATES

Year	Ex Post Depreciation	Ex Ante Depreciation	NIPA Depreciation
1950	−.5	27.5	5.7
1951	11.1	29.2	5.8
1952	7.5	22.4	5.6
1953	18.7	19.3	5.7
1954	−27.2	23.2	5.8
1955	−4.7	16.0	6.0
1956	8.1	12.6	6.3
1957	27.0	11.7	6.2
1958	−23.1	13.6	6.1
1959	7.8	10.1	6.2
1960	10.2	9.0	6.1
1961	−9.5	9.1	6.1
1962	17.5	8.8	6.1
1963	−4.9	10.2	6.0
1964	7	9.3	6.1
1965	3.2	9.6	6.2
1966	21.1	9.1	6.3
1967	−10.4	10.4	6.2
1968	.9	8.1	6.3
1969	25.4	5.8	6.4
1970	11.1	6.7	6.5
1971	−1.5	9.4	6.5
1972	2.4	8.4	
1973	31.9	6.2	
1974	43.7	8.5	
1975	−2.6	19.7	
1976	2.8	14.6	
1977	23.1	14.0	
1978	21.6	15.2	
1979	14.5	13.7	
1950–59	2.5	18.6	5.9
1960–69	5.3	8.9	6.2
1970–79	14.7	11.6	6.8
1950–79	7.5	13.0	6.3

future depreciation deductions, B_t, minus the real interest rate. C_t was calculated using data from the Economic Report of the President (Council of Economic Advisers 1982). The formula used was corporate profits plus capital consumption allowance with capital consumption adjustment, plus net interest, minus corporate profits taxes, minus 0.48 times the sum of NIPA capital consumption allowance and the capital consumption adjustment (the latter usually a negative number).

B was computed by taking B/p_K from Summers (1981) and multiplying by the sum of the current cost value of inventories plus property, plant, and equipment from the national balance sheets. The ex ante depreciation rate was calculated as $C_t/(MV_t - B_t)$ less the riskless interest rate, calculated by subtracting the inflation rate from the average 3-month Treasury bill rate.

Finally, NIPA depreciation rates were calculated by dividing NIPA depreciation by the current cost value of all nonfinancial corporate tangible assets at the end of the year, as reported in the national balance sheets.

In the table we present some data on depreciation at the aggregate level. Since the estimates are derived from market data, they pertain to all the assets of the corporate sector, not only those normally treated as depreciable. These include land and inventories, which are normally treated as having a zero depreciation rate, and account for about a third of nonfinancial corporate tangible assets based on information in the national balance sheets. This means that the NIPA depreciation figures we report are roughly a third lower than the composite rate on equipment and structures.

Broadly, the results corroborate the calculations presented so far. The mean ex post depreciation rate is 7.5, which is quite close to the standard estimates of the rate of depreciation. It is extremely volatile, ranging from 43 7 to −27.2. The mean ex ante depreciation rate is 13.0 reflecting its inclusion of the risk premium.

Including the recent Hulten and Wykoff (1981) used-asset price depreciation estimates, we now have four alternative measures of depreciation. The highest estimate is ex ante depreciation of 13.0 percent. This number is much higher than the ex post rate of 7.5 percent. By contrast, both the NIPA and Hulten and Wykoff estimates are significantly lower than the ex post rate. The NIPA rate is 6.3 percent, while Hulten and Wykoff are roughly 20 percent below the NIPA estimates for 1949–74.

The 5.5 percentage point differential between ex ante and ex post depreciation is remarkably in line with our prediction. Recall that with a risk premium on the market of 9.0 percent and a debt-equity ratio in the corporate sector of ½ we predict a 6.0 percent differential. Of course, estimates of the market risk premium include data from 1950–79. Nevertheless, it is worthy of note that two alternative empirical methodologies of estimating the difference between ex ante and ex post depreciation give virtually the same answer.

The most striking feature of the data is the increase in the ex ante rate of depreciation during the last decade. It averaged 7.8 between 1970 and 1974 and 15.4 between 1975 and 1979. This increase reflects the increased relative price uncertainty in the economy in three ways: first, the increased risk may have led to increased expected returns and higher risk premia. Second, increased relative price uncertainty increases the expected decline in an asset's price because the investor is likely to have made a more costly deviation from the ex post optimal choice of asset production techniques. Therefore, even if investors were risk neutral they would require higher rental price ratios because of increased expected economic depreciation. Third, the increased uncertainty about relative prices could have led firms to invest in less durable assets—particularly in the period right after the 1973–74 oil crisis when, with new information coming in rapidly, firms were no doubt leery of committing money to durable irreversible projects. For an excellent analysis and discussion of these issues, see Bernanke (1983).

VI. Conclusions

This paper makes the argument that analyses of the effects of capital taxation must recognize the importance of fluctuations in capital asset prices. We argue that most of the risk borne by owners of corporate capital pertains, not to the current rentals that are hedged by the corporate income tax, but to changes in the relative price of these assets that are not hedged because the tax is levied on accounting income. This means that the tax takes a much larger fraction of the return than it takes of the risk on corporate investments.

We then analyze the role of tax depreciation recognizing that there is substantial volatility in the rates at which capital assets lose their value. The tax system as currently set up relies on ex ante depreciation allowances rather than actual ex post measures of dep
and so does not share in the associated risk. We show that in
tic environment the natural counterpart to economic dep
involves allowing ex ante depreciation at a faster rate than
pected decline in asset values. More precisely, it depreciat
ances are to compensate investors for the risks they bear, the
of the risk premium in the asset's expected return that is attrib
to "capital risk" (asset price fluctuation) must be added to expected depreciation.

Our empirical analysis reveals that this adjustment is of substantial importance. Using two alternative methodologies, both based on financial market data, we conclude that the appropriate rate of ex ante depreciation to allow in the U.S. nonfinancial corporate sector is approximately twice that implied by data in the National Income Accounts. This suggests that many previous analyses have significantly understated the burden of taxes on corporate capital. It also suggests the need for further work in order to assess possible nonneutralities between assets of different durability and with different risk characteristics.

The research in this paper could be extended in a number of directions. The current analysis has ignored considerations of personal income taxation and corporate financial policy. We have not yet attempted an analysis of optimal taxation in the presence of capital risk. Such an analysis would need to recognize that the prices of inelastically supplied assets are likely to be more volatile than the prices of more elastically supplied assets. We hope to follow the valuable work of Hulten and Wykoff (1981) in using data on used-asset prices in conjunction with rental price data to derive depreciation rates on different types of assets.

Our analysis also has implications for work on tax reform. The scheme we propose of allowing ex ante depreciation deductions at a rate that compensates investors for the risks they bear is one of a number of equivalent types of tax reform. Similar goals could be achieved by including capital gains and losses, as measured on the stock market, in the corporate income tax. Alternatively, a tax on the net worth of the corporate sector could be employed. In the framework considered here, these tax schemes would be very similar. In the context of richer models there would be important differences that seem worthy of study.

The analysis here also has important implications for research on investment. The rate of ex ante economic depreciation appears in standard expressions for the cost of capital. Our calculations show that the rate increased substantially during the 1970s. This suggests a possible explanation for the sharp decline in net investment that has been observed during the 1970s.

References

Bernanke, Ben S. "Irreversibility, Uncertainty, and Cyclical Investment." *Q.J.E.* 98 (February 1983): 85–106.

Board of Governors of the Federal Reserve System. "Balance Sheets for the U.S. Economy 1945–80." Washington: Board of Governors, 1981.

Bulow, Jeremy I., and Shoven, John B. "Inflation, Corporate Profits, and the Rate of Return to Capital." In *Inflation: Causes and Effects*, edited by Robert Hall. Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), 1982.

Council of Economic Advisers. *Economic Report of the President.* Washington: Government Printing Office, 1982.

Feldstein, Martin S. "The Effects of Taxation on Risk Taking." *J.P.E.* 71 (September/October 1969): 755–64.

Fullerton, Don, and Gordon, Roger H. "A Reexamination of Tax Distortions in General Equilibrium Models." In *Behavioral Simulation Methods in Tax Policy Analysis*, edited by Martin S. Feldstein. Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), 1983.

Gordon, Robert J. "Energy Efficiency, User Cost Changes, and the Measurement of Durable Goods Prices." Working Paper no. 408. Cambridge, Mass.: Nat. Bur. Econ. Res., November 1979.

Gordon, Roger H. "Taxation of Corporate Capital Income: Tax Revenues vs.

Tax Distortions." Working Paper no. 687. Cambridge, Mass.: Nat. Bur. Econ. Res., June 1981.

Gordon, Roger H., and Malkiel, Burton G. "Corporation Finance." In *How Taxes Affect Economic Behavior,* edited by Henry J. Aaron and Joseph A. Pechman. Washington: Brookings Inst., 1981.

Holland, Daniel, and Myers, Stewart C. "Trends in Corporate Profitability and Capital Costs." In *The Nation's Capital Needs: Three Studies,* edited by Robert Lindsay. New York: Comm. Econ. Development, 1979.

Hulten, Charles R., and Wykoff, Frank C. "The Measurement of Economic Depreciation." In *Depreciation, Inflation, and the Taxation of Income from Capital,* edited by Charles R. Hulten. Washington: Urban Inst., 1981.

Ibbotson, Roger G., and Sinquefield, Rex A. *Stocks, Bonds, Bills, and Inflation: Historical Returns (1926–78).* Charlottesville, Va.: Financial Analysts Res. Foundation, 1979.

Jorgenson, Dale W., and Sullivan, Martin A. "Inflation and Corporate Capital Recovery." In *Depreciation, Inflation, and the Taxation of Income from Capital,* edited by Charles R. Hulten. Washington: Urban Inst., 1981.

Mintz, Jack M. "Some Additional Results on Investment, Risk Taking, and Full Loss Offset Corporate Taxation with Interest Deductibility." *Q.J.E.* 96 (November 1981): 631–42.

Stiglitz, Joseph E. "The Effects of Income, Wealth, and Capital Gains Taxation on Risk-Taking." *Q.J.E.* 83 (May 1969): 263–83.

Summers, Lawrence H. "Taxation and Corporate Investment: A *q*-Theory Approach." *Brookings Papers Econ. Activity,* no. 1 (1981), pp. 67–127.

———. "The Non-Adjustment of Nominal Interest Rates: A Study of the Fisher Effect." In *Macroeconomics, Prices, and Quantities: Essays in Memory of Arthur M. Okun,* edited by James Tobin. Washington: Brookings Inst., 1983.

Tobin, James. "Liquidity Preference as Behavior Towards Risk." *R. Studies* 25 (February 1958): 65–86.

A Monetary Equilibrium Model with Transactions Costs

Julio J. Rotemberg

Massachusetts Institute of Technology

This paper presents the competitive equilibrium of an economy in which people hold money for transactions purposes. It studies both the steady states that result from different rates of monetary expansion and the effects of such non-steady-state events as an open-market operation. Even though the model features no uncertainty and perfect foresight, open-market operations affect aggregate output. In particular, a simultaneous increase in money and governmental holdings of capital temporarily raises aggregate capital and output while it lowers the real rate of interest on capital.

I. Introduction

The objective of this paper is to study the competitive equilibrium of an economy in which people hold money for transactions purposes. As in the models of Baumol (1952), Tobin (1956), Stockman (1981), Jovanovic (1982), and Townsend (1982), but in contrast to those of Grandmont and Younes (1973), Lucas (1980), and Helpman (1981), households are allowed to hold interest-bearing capital in addition to barren money. The main advantage of the present model is that it is

---

I wish to thank Stanley Fischer, Timothy Kehoe, Robert Pindyck, Paul Romer, Robert Townsend, and especially Sanford Grossman for helpful comments and suggestions while retaining full responsibility for any remaining errors. I am also grateful to the National Science Foundation for its support through grant no. SES-8209266. This paper and those written independently by Grossman and Weiss (1982) and Grossman (1982) contain a very similar argument. However, the three papers differ in their details and should be viewed as complementary. Some of these differences are sketched in footnotes 2 and 6

[*Journal of Political Economy*, 1984, vol. 92, no. 1]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0002$01.50

able to shed light on the effects of such non-steady-state events as open-market operations.

Households pick the path of consumption optimally from their point of view. Because it is costly to carry out financial transactions, people visit their financial intermediary only occasionally. However, I do not let households pick optimally the length of the period during which they do not visit their bank. For tractability, the assumption is made that households have a constant interval during which they carry out no financial transactions.

A crucial feature of this paper is that, as in all free-market economies, different households visit their banks at different times. This leads to conclusions that are strikingly "Keynesian." Government interventions, and in particular open-market operations, have the ability to affect aggregate output and the real rate of interest. This is true even though the model features no uncertainty, full information, perfect foresight, and perfectly competitive markets for goods and money. The central finding of this paper is that a one-period monetary expansion leads to a lower real interest rate and a higher level of output for some time.

The paper proceeds as follows: Section II presents the model. It shows the maximization problems of the households and the firms, as well as the institutional environment. Section III presents the perfect-foresight equilibrium of the economy. It is a difference equation that, under certain conditions, exhibits saddle-path stability near the steady state that has positive consumption. Section IV studies steady-state inflation and welfare. It establishes that the level of output is indepe
dent of the rate of inflation but that inflation affects welf
tively by distorting the intertemporal consumption decisio
V is the heart of the paper since it discusses monetary poli
the steady state. Finally, Section VI presents some conclusi

II. The Model

There is only one good that can be both consumed and invested. Total output produced by competitive firms in period t (Q_t) depends, via a constant-returns-to-scale production function, on the amount of labor hired at t and on K_{t-1}, the amount of the good that was produced but not consumed at $t - 1$. Since an amount of labor $\bar{L}$ is assumed to be supplied inelastically,

$$Q_t = \bar{L}f\left(\frac{K_{t-1}}{\bar{L}}\right). \tag{1}$$

Accession Number.
84431
Date 19 · 3 · 86

where f is an increasing and concave function. For notational simplicity Q_t is total output and includes the depreciated value of K_{t-1}. Workers are assumed to be paid their marginal product. Therefore, the total return, denominated in period t goods, from forgoing the consumption of an additional good at $t - 1$ is given by $1 + r_{t-1}$, where[1]

$$1 + r_{t-1} = f'\left(\frac{K_{t-1}}{\bar{L}}\right). \tag{2}$$

There are $2n$ households. At time t, the households are assumed to maximize the utility function given by

$$V_t = \sum_{\tau=t}^{\infty} \rho^{\tau-t} \ln C_\tau^i, \tag{3}$$

where C_τ^i is the consumption of household i at time τ and ρ is a discount factor. The households have access to two assets, money and claims on capital. Money is the only asset with which goods can be bought. Moreover, visits to the financial intermediary for the purpose of converting claims on capital into money are costly. Therefore, as in the inventory theoretical models of Baumol (1952) and Tobin (1956), households engage in these visits only sporadically. In this paper it will be assumed at the outset that households exchange capital for money every two periods. The assumption that households do not change the timing of their financial transactions in response to events is made mainly for tractability. Except in stationary environments, it is very difficult to solve for the optimal timing of these visits, particularly when households pick their consumption path optimally.

Without loss of generality, suppose that household i engages in financial transactions in the "even" periods, t, $t + 2$, $t + 4$, ... At these dates it withdraws an amount M_τ^i of money balances that must be sufficient to pay for its consumption at τ and $\tau + 1$:

$$M_\tau^i = P_\tau C_\tau^i + P_{\tau+1} C_{\tau+1}^i, \tag{4}$$

where P_τ is the nominal price of the consumption and investment good at τ. Note that M_τ^i units of money are withdrawn at the beginning of period τ and are thus available to make purchases at τ.[2]

---

[1] Primes denote first derivatives, while double primes denote second derivatives.

[2] Grossman (1982) and Grossman and Weiss (1982) assume instead that money is withdrawn at the end of the period. Thus it is only available for consumption the following period.

The lifetime budget constraint of the household at t is given in terms of the household's monetary withdrawals:

$$\sum_{\tau=0}^{\infty} \frac{M^i_{t+2\tau}}{P_{t+2\tau}} \Big/ \prod_{i=0}^{2\tau-1} (1 + r_{t+i}) = \left[\sum_{\tau=0}^{\infty} Y^i_{t+\tau} \Big/ \prod_{i=0}^{\tau-1} (1 + r_{t+i})\right] + K^i_{t-1}(1 + r_{t-1}) - \left[\sum_{\tau=0}^{\infty} B \Big/ \prod_{i=0}^{2\tau-1} (1 + r_{t+i})\right], \quad (5)$$

where K^i_t are the claims on capital of household i at t, B is the real cost of visiting the financial intermediary, and Y^i_τ is the noninvestment income of the household at τ. This noninvestment income includes labor income as well as taxes and transfers from the government. Note that (5) explicitly assumes that noncapital income is directly invested in claims on capital. This assumption considerably simplifies the analysis.

The optimal path of consumption is found in two steps. First, I derive consumption at τ and $\tau + 1$ as a function of M^i_τ. Then I derive the optimal values of the sequence of monetary withdrawals.

The first step requires the maximization of

$$\ln C^i_\tau + \rho \ln C^i_{\tau+1}$$

subject to (4). This yields[3]

$$C^i_{\tau+1} = \frac{\rho P_\tau}{P_{\tau+1}} C^i_\tau = \frac{\rho}{1 + \rho} \frac{M^i_\tau}{P_{\tau+1}}.$$

Using (7), the expression in (6) is given by

$$\ln C^i_\tau + \rho \ln C^i_{\tau+1} = (1 + \rho) \ln\left(\frac{M^i_\tau}{P_\tau}\right) + \rho \ln \rho - (1 + \rho) \ln (1 + \rho) - \rho \ln\left(\frac{P_{t+1}}{P_t}\right). \quad (8)$$

Equation (8) asserts that the appropriately weighted sum of the instantaneous utilities at τ and $\tau + 1$ increases with M^i_τ/P_τ but is negatively affected by inflation between τ and $\tau + 1$.

---

[3] The analysis assumes that the holdings of money by household i from τ to $\tau + 1$ are positive. Here this is guaranteed by the fact that nonpositive holdings of money would induce nonpositive consumption and hence utility equal to minus infinity. However, if labor income were paid in the form of money, the constraint that monetary holdings be nonnegative might become binding.

Substituting (8) into (3) and using (5), one obtains

$$V_t = \sum_{k=0}^{\infty} \rho^{2k}\left[(1 + \rho) \ln \frac{M^t_{t+k}}{P_{t+k}} + \rho \ln \rho - (1 + \rho) \ln (1 + \rho) - \rho \ln \frac{P_{t+k+1}}{P_{t+k}}\right]. \tag{9}$$

This expression must now be maximized with respect to the sequence of monetary withdrawals subject to the lifetime budget constraint (5). This maximization yields

$$-\frac{1}{M^t_{t+2k}/P_{t+2k}} + \frac{\rho^2(1 + r_{t+2k})(1 + r_{t+2k+1})}{M^t_{t+2k+2}/P_{t+2k+2}} = 0, \tag{10}$$

as well as the transversality condition found, for instance, in Arrow and Kurz (1970):

$$\lim_{T\to\infty} \frac{\rho^T K^t_T}{M^t_T/P_T} = 0. \tag{11}$$

Using (7), (10) becomes

$$\frac{C^t_{t+2k+2}}{C^t_{t+2k}} = \rho^2(1 + r_{t+2k})(1 + r_{t+2k+1}). \tag{12}$$

Note that both (12) and (7) state that the marginal rate of substitution times a rate of return is equal to one. The important distinction between the two is that in (7) the rate of return is the rate of return on money, while in (12) it is the rate of return on capital. Stochastic versions of (12) have been statistically rejected using aggregate U.S. data by Mankiw (1981) and Hansen and Singleton (1982). Their rejections may be due in part to their neglect of the fact that in the presence of the transactions motive for holding money, the ratio of marginal utilities of consumption separated by different time intervals is related to rates of return of assets with different characteristics

The government in this model has no expenditures. However, it levies taxes, issues money, and holds capital. The evolution of the capital held by the government is given by[1]

$$K^G_{\tau+1} = f'(K_\tau)K^G_\tau + \frac{M_{\tau+1} - M_\tau}{P_{\tau+1}} + T_{\tau+1}, \tag{13}$$

where $T_{\tau+1}$ is the real taxes levied at $\tau + 1$, K^G_τ is the government's real holdings of capital at τ, and M_τ is high-powered money at τ. An increase in $M_{\tau+1}$ relative to M_τ will be called an open-market purchase and therefore the domain of monetary policy. Instead, a simultaneous change of $T_{\tau+1}$ and $K^G_{\tau+1}$ will be considered a type of fiscal policy. The government also requires that between periods τ and $\tau + 1$ (overnight) the monetary assets and liabilities of the financial intermediaries be equal to total high-powered money, M_τ. This is a form of a 100 percent reserve requirement in which the only form taken by high-powered money is deposits of intermediaries at the central bank. Hence, individuals' money consists of claims on the intermediaries only.

---

[1] For simplicity, I ignore the transaction costs incurred by the government when it engages in an open-market operation

Financial intermediaries receive the household's income and invest it in claims on capital. They are also allowed to issue a certain quantity of money. The intermediaries are compensated for their services with the brokerage fees, B, of (5). Their function can best be understood by following their transactions in detail.

Between periods τ and $\tau + 1$ the financial intermediaries have as their assets the household's claims on capital as well as M_τ units of money. Their liabilities are the household's claims on capital and the monetary assets of the households. The households that visit the intermediaries at $\tau + 1$ do not carry over any money between τ and $\tau + 1$ if the nominal rate of interest is positive, which is assumed throughout. So the monetary liabilities of the intermediaries between τ and $\tau + 1$ consist solely of the amount withdrawn but not spent by the households that visit the intermediaries at τ. These liabilities cannot, by law, exceed M_τ. The intermediaries would, of course, like to issue more money as long as the nominal rate of interest is positive

At the beginning of $\tau + 1$, the financial intermediaries is- [illegible] ever amount of money ($nM^l_{\tau+1}$, where n is the number of ho [illegible] is required by the households that visit them at $\tau + 1$. M [illegible] their nonmonetary assets and liabilities have grown since [illegible] now must pay interest on their debt as well as pay the ho [illegible] their labor income. Both these payments are made through [illegible] mediaries. Therefore, the intermediaries' consolidated balance sh [illegible] measured in nominal dollars is given by table 1, where K^P_τ and Y^l_τ are the amount of capital owned by the private sector and labor income at τ, respectively.

TABLE 1

CONSOLIDATED BALANCE SHEET OF THE INTERMEDIARIES AT THE BEGINNING OF $\tau + 1$

	Assets	Liabilities
Nonmonetary	$\left[K^P_\tau f'\left(\frac{K_\tau}{L}\right) + Y^l_{\tau+1}\right]P_{\tau+1}$	$\left[K^P_\tau f'\left(\frac{K_\tau}{L}\right) + Y^l_{\tau+1}\right]P_{\tau+1} - nM_{\tau+1}$
Monetary	M_τ	$nM^l_{\tau+1} + M_\tau$

During the period $\tau + 1$, the households make purchases whose total value is $P_{\tau+1}C_{\tau+1}$, where $C_{\tau+1}$ is total consumption at $\tau + 1$. These purchases are paid with money that the firms return to the intermediaries at the end of the period in partial payment of their obligations. The firms also sell, for money that they likewise return to the intermediaries, $(M_{\tau+1} - M_\tau)/P_{\tau+1}$ units of capital to the government. Finally, the government levies taxes directly on the nonmonetary assets held by the households at the intermediaries. Therefore the consolidated balance sheet of the intermediaries at the end of $\tau + 1$ is given by table 2.

The government requires that the monetary liabilities be equal to $M_{\tau+1}$. Thus, $nM'_{\tau+1} = P_{\tau+1}C_{\tau+1} + M_{\tau+1} - M_\tau$. Substituting for $nM'_{\tau+1}$ in the nonmonetary liabilities, it is clear that the nonmonetary liabilities equal the nonmonetary assets. Moreover,

$$\begin{aligned} P_{\tau+1}\left[K_\tau^P f'\left(\frac{K_\tau}{\bar{L}}\right) + Y^L_{\tau+1} - C_{\tau+1} - T_{\tau+1}\right] - (M_{\tau+1} - M_\tau) \\ = P_{\tau+1}\left[(K_\tau^P + K_\tau^G) f'\left(\frac{K_\tau}{\bar{L}}\right) + Y^L_{\tau+1} - K^G_{\tau+1} - C_{\tau+1}\right] \quad (14) \\ = P_{\tau+1}(Q_{\tau+1} - C_{\tau+1} - K^G_{\tau+1}) = P_{\tau+1}K^P_{\tau+1}, \end{aligned}$$

where the first equality is obtained using (13). The second equality follows from the fact that $K_\tau^P + K_\tau^G = K_\tau$ and that, under constant returns to scale, the value of output is equal to factor payments. Finally, the last equality results from using the identity that makes output, Q_τ, equal to consumption, C_τ, plus investment, K_τ. Hence, again, the nonmonetary assets and liabilities of the intermediaries between $\tau + 1$ and $\tau + 2$ are equal to the private holdings of capital.

I also assume that half the households (n) visit the intermediaries in the even periods $t, t + 2, t + 4, \ldots$, and the other half carry out their financial transactions in the odd periods. The fact that only a subset of the households visit financial intermediaries in any given day is one of the main features of reality that this paper seeks to reflect. It also is a feature of the steady states studied by Jovanovic (1982). It turns out that the assumption that households stagger their financial transactions is crucial to ensure that open-market operations have real effects.

III. Equilibrium

The equilibrium for this economy is a path for the price level and for the real rate of interest such that households maximize utility and firms maximize profits using these prices and the following two conditions are met.

TABLE 2

CONSOLIDATED BALANCE SHEET OF THE INTERMEDIARIES AT THE END OF $\tau + 1$

	Assets	Liabilities
Nonmonetary	$P_{\tau+1}\left[K_\tau^P f'\left(\frac{K_\tau}{\bar{L}}\right) + Y^l_{\tau+1} - C_{\tau+1} - T_{\tau+1}\right] - (M_{\tau+1} - M_\tau)$	$P_{\tau+1}\left[K_\tau^P f'\left(\frac{K_\tau}{\bar{L}}\right) + Y^l_{\tau+1} - T_{\tau+1}\right] - nM'_{\tau+1}$
Monetary	$M_\tau + (M_{\tau+1} - M_\tau)$	$nM'_{\tau+1} + M_\tau - P_{\tau+1}C_{\tau+1}$

1. The sum of consumption and capital demanded at τ by the households and capital demanded by the government at τ is equal to output at τ:

$$C_\tau + K_\tau^P + K_\tau^G = \bar{L}f\left(\frac{K_{\tau-1}}{\bar{L}}\right). \tag{15}$$

2. The amount of money that households that visit the intermediaries at τ want to hold between τ and $\tau + 1$ must be equal to M_τ. Hence, the total expenditures at $\tau + 1$ by households that visit the financial intermediaries at τ must be equal to M_τ.

Let C_τ^τ and $C_\tau^{\tau-1}$ be the consumptions at τ of households that visit the financial intermediary at τ and $\tau - 1$, respectively. Then condition (2) requires that

$$nP_\tau C_\tau^{\tau-1} = M_{\tau-1}. \tag{16}$$

Therefore, using (7),

$$n\rho P_\tau C_\tau^\tau = M_\tau. \tag{17}$$

Therefore,

$$C_\tau^{\tau-1} = \frac{\rho M_{\tau-1}}{M_\tau} C_\tau^\tau, \tag{18}$$

and aggregate consumption at τ, C_τ, is given by

$$C_\tau = n\left(1 + \frac{\rho M_{\tau-1}}{M_\tau}\right)C_\tau^\tau. \tag{19}$$

Using (19), (11), and the equilibrium condition (15), one obtains the difference equation that governs the evolution of aggregate capital:

$$\bar{L}f\left(\frac{K_{\tau+2}}{\bar{L}}\right) - K_{\tau+3} = \rho^2 \frac{1 + \rho(M_{\tau+2}/M_{\tau+3})}{1 + \rho(M_\tau/M_{\tau+1})} f'\left(\frac{K_{\tau+1}}{\bar{L}}\right) f'\left(\frac{K_{\tau+2}}{\bar{L}}\right)$$

$$\times \left[\bar{L}f\left(\frac{K_\tau}{\bar{L}}\right) - K_{\tau+1}\right], \tag{20}$$

$$\tau = t - 1, t, t + 1, \ldots .$$

Knowledge of the sequence of capitals provides the sequence of rates of return by (2), the aggregate consumptions by (15), the sequence of individual consumptions by (19), and the sequence of prices by (17). The equilibrium is thus a third-order nonlinear difference equation with, so far, only one initial condition, namely, K_{t-1}. It will be shown below that equation (20), which is a necessary condition for an equilibrium, establishes that open-market operations are nonneutral. Before studying this question, however, I focus on the existence and uniqueness of an equilibrium that converges to a steady state with positive consumption. As long as $M_\tau/M_{\tau+1}$ converges to a constant, the steady-state values of capital (K^*) have the following property:

$$\left[\bar{L}f\left(\frac{K^*}{\bar{L}}\right) - K^*\right]\left\{1 - \rho^2\left[f'\left(\frac{K^*}{\bar{L}}\right)\right]^2\right\} = 0.$$

The steady-state values of K do not depend on the rate of monetary growth. There are two types of steady states. Those with zero consumption are such that output $\bar{L}f(K^*/\bar{L})$ is equal to investment, K^*. These are ruled out by the transversality condition (11). The only steady state with positive consumption has the property that $\rho f'(\bar{K}) = 1$; the product of the discount rate times the marginal product of capital is unity. Unfortunately, I cannot establish the existence or uniqueness of paths that converge to $\bar{K}$. I can only present a local result, namely, that the linearized version around $\bar{K}$ of (20) is such that, under certain conditions, a unique path that converges to $\bar{K}$ exists. The linearized version of (20) around $\bar{K}$, assuming money growth is constant, is

$$\begin{aligned}(K_{\tau+3} - \bar{K}) &- \left\{f'(\bar{K}) - \rho\left[f\left(\frac{\bar{K}}{\bar{L}}\right) - \frac{\bar{K}}{\bar{L}}\right]f''\left(\frac{\bar{K}}{\bar{L}}\right)\right\}(K_{\tau+2} - \bar{K}) \\ &- \left\{1 - \rho\left[f\left(\frac{\bar{K}}{\bar{L}}\right) - \frac{\bar{K}}{\bar{L}}\right]f''\left(\frac{\bar{K}}{\bar{L}}\right)\right\}(K_{\tau+1} - \bar{K}) + f'\left(\frac{\bar{K}}{\bar{L}}\right)(K_\tau - \bar{K}) = 0\end{aligned} \tag{21}$$

or $(1 - \lambda_1 L)(1 - \lambda_2 L)(1 - \lambda_3 L)(K_\tau - \bar{K}) = 0$, where L operator and the roots λ_1, λ_2, and λ_3 have the following p

$$\lambda_1\lambda_2\lambda_3 = -f'\left(\frac{\bar{K}}{\bar{L}}\right),$$

$$\lambda_1\lambda_2 + \lambda_2\lambda_3 + \lambda_1\lambda_3 = -1 + \rho\left[f\left(\frac{\bar{K}}{\bar{L}}\right) - \frac{\bar{K}}{\bar{L}}\right]f''\left(\frac{\bar{K}}{\bar{L}}\right).$$

$$\lambda_1 + \lambda_2 + \lambda_3 = f'\left(\frac{\bar{K}}{\bar{L}}\right) - \rho\left[f\left(\frac{\bar{K}}{\bar{L}}\right) - \frac{\bar{K}}{\bar{L}}\right]f''\left(\frac{\bar{K}}{\bar{L}}\right)$$

Inspection of these equations reveals that one of the roots, say λ_1, is equal to minus one,[5] while the other two roots are such that $(\lambda_2 \; 1)(\lambda_3 - 1)$ is negative, $\lambda_2\lambda_3$ is positive, and $(\lambda_2 + \lambda_3)$ is positive. Therefore, λ_2 and λ_3 are both positive and lie on opposite sides of the unit circle. There is one stable root, say λ_3. This root corresponds to the initial condition that gives, at t, the historical level of K_{t-1}. The root bigger than one, λ_2, is such that, if the initial conditions are not chosen correctly, K either explodes or implodes, contradicting the transversality condition (11). The third root, λ_1, must ensure that the path is feasible not only for the economy as a whole but also for the two types of households. This initial condition can be obtained as follows. Using (10) and (7) for a household that withdraws money at t, (5) becomes

---

[5] To see that one of the roots is equal to minus one, it suffices to check that in (21) one (the coefficient of $K_{\tau+3}$) minus the coefficient of $K_{\tau+2}$ plus the coefficient of $K_{\tau+1}$ minus the coefficient of K_τ is equal to zero

$$\begin{aligned}\frac{(1+\rho)}{1-\rho^2}C^t_t = {} & \left[\sum_{\tau=0}^{\infty} Y^t_{t+\tau} \Big/ \prod_{i=0}^{\tau-1}(1+r_{t+i})\right] + K^t_{t-1}(1+r_{t-1}) \\ & - \left[\sum_{\tau=0}^{\infty} B \Big/ \prod_{i=0}^{2\tau-1}(1+r_{t+i})\right],\end{aligned} \tag{22}$$

where $Y^t_{t+\tau}$ and K^t_{t-1} are the noncapital income earned at $t+\tau$ and the capital holdings at $t-1$ of an agent who visits the intermediaries at t. The sequence of Y's in (22) is endogenous, so (22) is not an appropriate initial condition. Instead, consider the difference between C^{t+1}_{t+1} and $C^t_t(1+r_t)$, where C^t_t is given by (22) and C^{t+1}_{t+1} is given by the corresponding equation for the individual who goes to the intermediaries at $t+1$:

$$\begin{aligned}[C^{t+1}_{t+1} - C^t_t(1+r_t)]\frac{(1+\rho)}{1-\rho^2} = {} & (Y^{t+1}_t - Y^t_t)(1+r_t) \\ & + \left[\sum_{\tau=0}^{\infty}(Y^{t+1}_{t+1+\tau} - Y^t_{t+1+\tau}) \Big/ \prod_{i=0}^{\tau-1}(1+r_{t+i+1})\right] \\ & + (K^{t+1}_t - K^t_{t-1})(1+r_{t-1})(1+r_t) + B(1+r_t) \\ & + \left[\sum_{\tau=1}^{\infty} B \Big/ \prod_{i=1}^{2\tau-1}(1+r_{t+i})\right] - \left[\sum_{\tau=0}^{\infty} B \Big/ \prod_{i=0}^{2\tau-1}(1+r_{t+i+1})\right].\end{aligned}$$

Assuming Y^t_τ is equal to Y^{t+1}_τ, ignoring the transaction costs B, and using (19) this equation becomes

$$\begin{aligned}\frac{1+\rho}{n}\left\{\frac{\bar{L}f(K_t/\bar{L}) - K_{t+1}}{1+\rho(M_t/M_{t+1})} - \frac{f'(K_t/\bar{L})[\bar{L}f(K_{t-1}/\bar{L}) - K_t]}{1+\rho(M_{t-1}/M_t)}\right\} \\ = (K^{t+1}_t - K^t_{t-1})f'\left(\frac{K_{t-1}}{\bar{L}}\right)f'\left(\frac{K_t}{\bar{L}}\right)(1-\rho^2),\end{aligned} \tag{23}$$

which is another initial condition.

Unfortunately, as pointed out by Samuelson (1947), for instance, the presence of a root with modulus one in the linearized equation is consistent with both stability and instability of the nonlinear equation.

The higher-order terms matter. The oscillatory root λ_1 could "correspond" to convergent, oscillatory, or divergent behavior. In other words, the stable manifold of (20) can be either two- or one-dimensional. In the former case, the initial condition given by (23) ensures that there is a unique equilibrium that converges to $\bar{K}$.[6] In the latter case, there may exist no convergent equilibrium that satisfies (23).[7] In other words, the path that converges to $\bar{K}$ may not satisfy (23). However, this convergent path will still be an equilibrium if the government carries out lump-sum transfers that ensure that, eventually, the wealth of the two types of agents is the same so that consumption is constant.

IV. Inflation, Steady States, and Welfare

In Section III, I established that there is no Tobin effect in this model. Independently of the value of M_{t+1}/M_t, the unique steady-state value of the capital stock that involves positive consumption is $\bar{K}$. This does not mean that inflation has no real effects. In particular, the steady-state rate of growth of the money stock affects the path of individual consumption, the level of welfare, and the income velocity of money.

Before studying the effects of inflation, however, it must be ascertained that an inflationary path is consistent with the government's budget constraint (14). I will assume that the government lets the money stock grow at the rate m so that $M_\tau/M_{\tau-1} = 1 + m$. With the new money, the government buys capital, which it redistributes in lump-sum fashion. These lump-sum redistributions affect none of the conditions used to derive (20). So the amount of capital l
the government can be arbitrarily set to some constant.

I now compute the rate of inflation, which corresponds t
rate of growth of money. Equation (17) establishes that th
state rate of inflation, π, is given by

$$\frac{P_{\tau+1}}{P_\tau} = 1 + \pi = \rho \frac{C^\tau_\tau}{C^{\tau-1}_\tau},$$

where in the steady state neither C^τ_τ nor $C^{\tau-1}_\tau$ depends on τ. Therefore, using (18),

$$1 + \pi = \frac{M_\tau}{M_{\tau-1}} = 1 + m. \qquad (25)$$

---

[6] Grossman (1982) and Grossman and Weiss (1982) consider a situation with no capital and in which output is constant. In their model, the oscillatory root is always convergent.

[7] The properties of equilibria in this case are unknown. However, it is possible that the equilibrium settles down to an oscillation between two levels of capital $\check{K}$ and $\hat{K}$ such that $\rho^2 f'(\check{K}) f'(\hat{K})$ is equal to one. Such oscillations satisfy (20) as long as consumption oscillates between $\bar{L}f(\hat{K}/\bar{L}) - \check{K}$ and $\bar{L}f(\check{K}/\bar{L}) - \hat{K}$.

The rate of inflation is equal to the rate of monetary expansion. Note that the model is quite consistent with the rate-of-return dominance of capital over money in the steady state. The rate of return of the former is $(1/\rho) - 1$, while that of the latter is $-m/(1 + m)$.

By (24), the ratio of consumption on the date of financial transactions to consumption in the following period is $(1 + m)/\rho$; it rises as inflation rises. Therefore, inflation distorts the intertemporal consumption decisions. It leads people to consume more right after they withdraw money and less in later periods.[8] The rate of deflation, which is such that intertemporal consumption decisions are optimal from the point of view of society, is $1 - \rho$. As Friedman (1969) proposed, the rate of deflation must be equal to the discount rate. This result was also obtained by Jovanovic (1982) in a model in which money is held for transactions purposes but in which people, while picking the timing of their visits to the financial intermediaries optimally, do not pick their consumption path optimally. Here it can be shown as follows. Consider a social planner who wants to maximize

$$\sum_{\tau=1}^{\infty} \rho^{\tau-1}[\alpha \ln C_\tau^\tau + (1-\alpha) \ln C_\tau^{\tau-1}] \tag{26}$$

subject to

$$K_\tau = \bar{L} f\left(\frac{K_{\tau-1}}{\bar{L}}\right) - n(C_\tau^\tau + C_\tau^{\tau-1}) - nB, \tag{27}$$

where α is a weight between zero and one. This social planner maximizes a convex combination of the utilities of both types of consumers subject to society's budget constraint. The plan that maximizes (26) satisfies $C_\tau^\tau = [\alpha/(1-\alpha)]C_\tau^{\tau-1}$ and

$$C_{\tau+1}^\tau = \rho f'\left(\frac{K_{\tau-1}}{\bar{L}}\right) C_\tau^\tau. \tag{28}$$

In the steady state in which $\rho f'(K_\tau/\bar{L})$ is equal to one, consumption is constant. Indeed, this is precisely what occurs in the decentralized economy with money as long as, in (7), $\rho P_t/P_{t+1}$ is equal to one. This in turn requires that m be equal to $(\rho - 1)$ as claimed.

However, there is a problem in sustaining the equilibrium with $K = \bar{K}$ when m is equal to $(\rho - 1)$. This problem arises because at this equilibrium money and capital have the same rate of return. Therefore households would prefer to withdraw money at the beginning of their lives in the amount equal to the present discounted value of their income. Then they would avoid all visits to their bank and associated transactions costs. This would result in no capital being available in this model with 100 percent reserves. It may well be the case that, with a smaller reserve requirement, the equilibrium would be sustainable. In any event, note that, as long as m is just slightly bigger than $(\rho - 1)$, the equilibrium of this model is essentially equal to the Pareto optimum and does induce people to hold $\bar{K}$ units of capital.

---

[8] This has also been noted by Jovanovic (1982). This effect is likely to be even more pronounced on consumer expenditure when there are durable goods, and consumer expenditure can be different from consumption. In this paper, consumption and consumer expenditure coincide by assumption.

Inflation induces people to consume less in the period in which they do not go to the bank. Therefore, $M_\tau/P_{\tau+1}$, which is equal to this consumption, falls. However, surprisingly, in this model M_τ/P_τ actually rises with inflation. This result is undoubtedly due in part to the fact that people do not go to the bank more often when inflation rises. It emerges because, even though people want to reduce $M_\tau/P_{\tau+1}$, they must increase M_τ/P_τ to ensure that inflation does not reduce $M_\tau/P_{\tau+1}$ too much. The result can be established by noting that equation (18) says that M_τ/P_τ is proportional to consumption in the period in which people visit the financial intermediary. By equation (19), this consumption does indeed rise with inflation.

V. The Nonneutrality of Monetary Policy

The main purpose of this paper is to study conditions outside th
steady state. First, it will be established that a wide variety of m
policies (or open-market operations) affect aggregate outpu
the consensus of textbooks, such as Branson (1979). How
view has recently been challenged by a variety of authors (
lace 1981; Chamley and Polemarchakis 1983). These auth
shown that in models in which money is held only for its rate
characteristics, open-market operations are neutral. Admit
premise of these models—that money is not rate of return d
by other assets—appears to be a bad description of free-market
omies as we know them. The proof that open-market operations can affect output in the economy of this paper is straightforward. Consider a base path for money and taxes, $\{\tilde{M}_\tau\}$ and $\{\tilde{T}_\tau\}$; then the equilibrium sequence of capital $\{\tilde{K}_\tau\}$ must satisfy (20).

Now consider a slightly different financial policy for the government. At t, unexpectedly, the government purchases some extra capital by issuing ϵ units of money. Then, at T, the government engages in the reverse transaction: it sells $\epsilon[\Pi_{\tau=t}^{T-1} f'(K_\tau/\bar{L})]/P_t$ units of capital. Therefore, the path of taxes remains unchanged, but between t and T the path of money is replaced by $\{M_\tau + \epsilon\}$. After T, the path of money is given by $\{\tilde{M}_\tau + \epsilon[1 - P_T \Pi_{\tau=t}^{T} f'(K_\tau/\bar{L})]/P_t\}$. Then, it is clear from (20) that if, in the new equilibrium, capital remains unchanged from

$\{\tilde{K}_\tau\}$ at t and $t + 1$, it will have to be different from $\tilde{K}_{t+2}$ at $t + 2$; monetary policy affects output.

The intuition behind this result is as follows. Suppose the open-market operation had no effect on prices. Then the paths of consumption would be unaffected. However, the people who visit the bank at t would not want to demand the increased amount of money balances. If, instead, any rate of inflation after t were affected by the open-market operation, then by (7) some individuals would change their consumption path. Finally, suppose that only the price at t was affected by the government's change of financial policy. Then the consumption of those who visited the bank at $t - 1$, which is given by M_{t-1}/P_t, would be affected. So the nonneutrality of open-market operations hinges crucially on the fact that not everyone visits the bank on the day of the operation.

I am not just interested in establishing that monetary policy is nonneutral. Instead I want to characterize its effect. To do this for general policies, it would be necessary to solve the nonlinear difference equation (20). Two simpler approaches are pursued here. The first is to simulate the unique nonexplosive path given by equation (20) and the initial condition (23). This nonexplosive path is obtained numerically by assuming that $\bar{L}$ is equal to one, the production function is given by $Q_t = K_t^{.75}$, and ρ is equal to .99. The steady state of this economy has a capital stock of .30394. I assume that in period 0 the economy is at this steady state. Therefore the difference in the holdings of capital of the two types of people at zero, $n(K_0^2 - K_0^1)$, is given by $\rho[\bar{K} - f(\bar{K})]/(1 + \rho)$. Figure 1 presents the path of capital that ensues from a 2 percent increase in the money stock that occurs unexpectedly in period 1. This path is obtained by searching over different values of K_1. For each value of K_1, (23) is used to obtain K_2 and (20) is used to obtain the latter levels of capital. Values of K_1 slightly lower than those reported lead to implosions in capital, while slightly higher ones lead to explosions. Interestingly, for the parameters I considered, the system does not tend to oscillate. Figure 1 shows that the monetary expansion increases capital, albeit by a small amount. Capital then slowly returns to its steady-state level. The intuition behind these results is the following. When the quantity of money is increased at τ, the price level rises. This decreases $M_{\tau-1}/P_\tau$ and therefore reduces $C_\tau^{\tau-1}$. This fall in the consumption of those who do not visit their bank at τ raises capital and hence output in the following period. This explanation suggests that monetary policy may derive much of its power in this model from the assumption that those people who visited the bank at $\tau - 1$ do not change their pattern of bank visits in response to inflation at τ. How much people who had

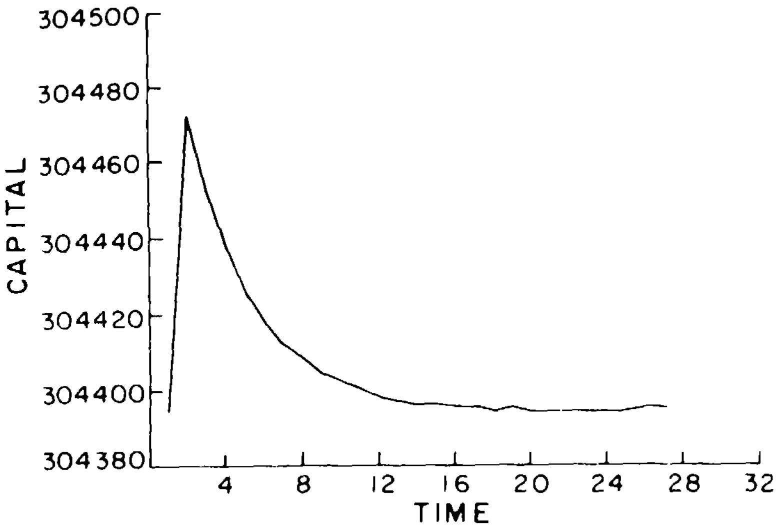


FIG. 1.—Open-market monetary increase in period 2

not scheduled a visit to their intermediary at τ would reduce their consumption in response to inflation at τ if they were free to pick the timing of these visits optimally is an open question that deserves further research.

Figure 2 shows the effects on the path of capital of a 2
increase in the money supply in period 5 that is announced
1. Capital and output immediately begin to fall and contin
until period 5. Then they both jump up. This jump is not sig
smaller than the one produced by an unexpected moneta
sion. The low level of capital in the period just preceding th
tary expansion can be explained as follows. The monetary exp
leads people in period 4 to expect high inflation. Therefore, the households that visit the intermediaries in period 4 consume a disproportionate amount at 4 and thus lower the capital stock. The fall in the capital stock the moment the future monetary increase is announced has an unexpected consequence. It means that output increases more when an unexpected monetary injection's effect on government capital is expected to be neutralized by a future open-market operation than when it is expected to be neutralized by giving equal lump-sum transfers to the entire population.

The second approach is to compute the solution using the linearized versions of (20) and (23) around the steady state. Assuming that $\bar{L} =$ 1 and that g_{t+3}, which is given by $[f(\bar{K}) - \bar{K}]\{[1 + \rho(M_{\tau+2}/$

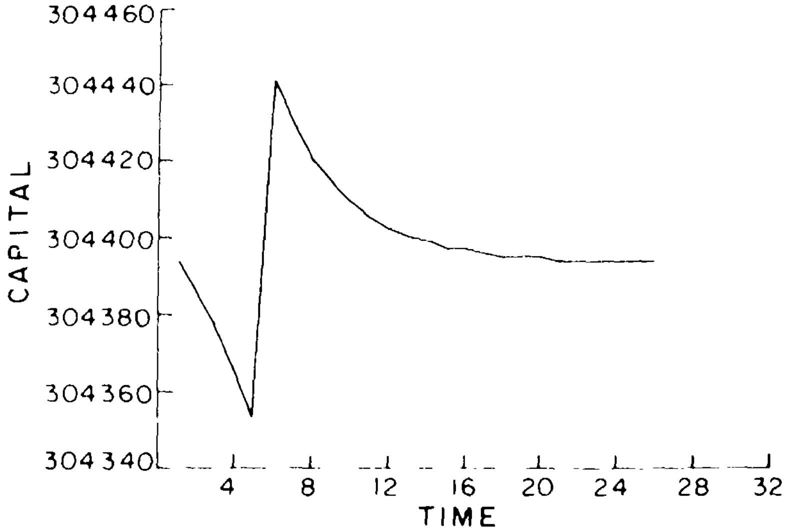


FIG. 2.—Open-market monetary increase in period 5

$M_{T+3})]/[1 + \rho(M_T/M_{T+1})] - 1\}$, is close to zero, the linearized version of (20) becomes

$$(1 + L)(1 - \lambda_2 L)(1 - \lambda_3 L)(K_{T+3} - \bar{K}) = -g_{T+3}. \qquad (29)$$

The nonexplosive solution of (29) is given by

$$\begin{aligned}(K_T - \bar{K}) &= \lambda_3(K_{T-1} - \bar{K}) + A(-1)^T \\ &\quad + \frac{1}{1 + \lambda_2}\sum_{j=0}^{\infty}\left[\left(\frac{1}{\lambda_2}\right)^j - (-1)^j\right]g_{T+j+1},\end{aligned} \qquad (30)$$

where A must be picked to satisfy the version of (23) that is linearized at the steady state[9]

$$\begin{aligned}(K_{t+1} - \bar{K}) &- \{2f'(\bar{K}) - \rho f''(\bar{K})[f(\bar{K}) - \bar{K}]\}(K_t - \bar{K}) \\ &+ \{f'(K)^2 - (1 - \rho)f''(\bar{K})[f(\bar{K}) - \bar{K}]\}(K_{t-1} - \bar{K}) \\ &= -f'(\bar{K})g_{t+2}.\end{aligned} \qquad (31)$$

Using (30) for t and $t + 1$, assuming t is even, and substituting in (31), one obtains A:

---

[9] For simplicity, (31) ignores money growth between t and $t + 2$.

$$
\begin{aligned}
A = {} & \left\{\frac{\lambda_3 + f'(\bar{K})^2 - (1-\rho)f''(\bar{K})[f(\bar{K}) - K]}{B} - \lambda_3\right\}(K_{t-1} - \bar{K}) \\
& + \left[\frac{f'(\bar{K})}{B} + \frac{1}{\lambda_2 B} - \frac{1}{\lambda_2}\right]g_{t+2} + \frac{1}{1+\lambda_2} \\
& \times \left[\sum_{j+1}^{\infty}\left(\frac{1}{\lambda_2}\right)^j\left(\frac{1}{B} + \frac{1}{\lambda_2 B} - \frac{1}{\lambda_2}\right)g_{t+j+2} - \sum_{j+1}^{\infty}(-1)^j g_{t+j+2}\right],
\end{aligned}
\tag{32}
$$

$$B = 1 - \lambda_3 + 2f'(\bar{K}) - \rho f''(\bar{K})[f(\bar{K}) - \bar{K}].$$

An unexpected increase in money at t raises g_{t+2} and leaves all g_τ's for τ greater than $t+2$ unaffected. So, it raises K_t by $[f'(\bar{K}) + (1/\lambda_1)]/B$. Assuming the same technology and monetary shock that lead to figure 1, the change in K_t is .0007838. This is very similar to the change of .000778 computed to arrive at figure 1. However, since, in the linear approximation, A is equal to .0000062, the linear approximation exhibits small oscillations in the steady state that are absent in the nonlinear system. If the linear approximation is forced to converge by choosing A equal to zero, the change in capital at t is .0007826. So, even though the presence of the unit root implies that the linearized system is not, as a general rule, a good approximation to the behavior of the nonlinear system, it appears to give reasonable answers in the context studied here.

VI. Conclusions

The model of this paper is a modest step toward the constru
tractable and realistic general equilibrium models capable
ding light on the effects of non-steady-state monetary chan
major advantage is that people's motive for holding money is exp
itly that money is used for transactions. In particular, in those periods in which households do not visit their banks, they are faced with an extreme version of the "Clower constraint": they must pay for their purchases with money carried over from the previous period. This ensures that monetary policies that expand money and prices reduce the real consumption of those households that do not visit their financial intermediary on the day of the monetary expansion. This fall in consumption raises capital and output in future periods.

A number of issues are raised by this paper. First, an important question is to what extent the power of monetary policy would be diluted if people timed their visits to intermediaries optimally. Associated with this question is the empirical question of whether people in fact do significantly alter the interval during which they refrain from visiting their bank as events change.

It is hoped that the framework of this paper can also be used to study the effects of various institutional setups on macroeconomic activity. In particular, it should be capable of shedding some light on the difference between commodity standards, fractional reserve standards, and the 100 percent reserves standard of this paper.

References

Arrow, Kenneth J., and Kurz, Mordecai. *Public Investment, the Rate of Return and Optimal Fiscal Policy.* Baltimore: Johns Hopkins Press (for Resources for the Future), 1970.

Baumol, William J. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *Q.J.E.* 66 (November 1952): 545–56.

Branson, William H. *Macroeconomic Theory and Policy.* 2d ed. New York: Harper & Row, 1979.

Chamley, Christophe, and Polemarchakis, Heraklis. "Asset Markets, Returns and the Neutrality of Money." *Rev. Econ. Studies* (1983), in press.

Friedman, Milton. *The Optimum Quantity of Money and Other Essays.* Chicago: Aldine, 1969.

Grandmont, Jean-Michel, and Younes, Yves. "On the Efficiency of a Monetary Equilibrium." *Rev. Econ. Studies* 40 (April 1973): 149–65.

Grossman, Sanford J. "A Transactions-based Model of the Monetary Transmission Mechanism: Part 2." Working Paper no. 974. Cambridge, Mass.: Nat. Bur. Econ. Res., September 1982.

Grossman, Sanford J., and Weiss, Laurence. "A Transactions-based Model of the Monetary Transmission Mechanism: Part I." Working Paper no. 973. Cambridge, Mass.: Nat. Bur. Econ. Res., September 1982.

Hansen, Lars Peter, and Singleton, Kenneth J. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." *Econometrica* 50 (September 1982): 1269–86.

Helpman, Elhanan. "Optimal Spending and Money Holdings in the Presence of Liquidity Constraints." *Econometrica* 49 (November 1981): 1559–70.

Jovanovic, Boyan. "Inflation and Welfare in the Steady State." *J.P.E.* 90 (June 1982): 561–77.

Lucas, Robert E., Jr. "Equilibrium in a Pure Currency Economy." *Econ. Inquiry* 18 (April 1980): 203–20.

Mankiw, N. Gregory. "The Permanent Income Hypothesis and the Real Interest Rate." *Econ. Letters* 7, no. 4 (1981): 307–11.

Samuelson, Paul A. *The Foundations of Economic Analysis.* Cambridge, Mass.: Harvard Univ. Press, 1947.

Stockman, Alan C. "Anticipated Inflation and the Capital Stock in a Cash-in-Advance Economy." *J. Monetary Econ.* 8 (November 1981): 387–93.

Tobin, James. "The Interest-Elasticity of Transactions Demand for Cash." *Rev. Econ. and Statis.* 38 (August 1956): 241–47.

Townsend, Robert M. "Asset Market Anomalies in a Monetary Economy." Mimeographed. Pittsburgh: Carnegie-Mellon Univ., 1982.

Wallace, Neil. "A Modigliani-Miller Theorem for Open-Market Operations." *A.E.R.* 71 (June 1981): 267–74.

Earnings Inequality among Males in the United States: Trends and the Effect of Labor Force Growth

Martin D. Dooley

McMaster University

Peter Gottschalk

Bowdoin College and Institute for Research on Poverty

We analyze the variance of log earnings within labor force cohorts of U.S. males with particular attention to the influence of labor force growth rates. Estimates with data from the 1968–79 current population surveys reveal increases in earnings inequality within labor force cohorts even after we control for the level of education, experience, and unemployment. A model of human capital investment is used to analyze differences among cohorts in the life-cycle profile of the variance of log earnings. It is shown that a pattern of sharply rising and then declining labor force growth rates will tend to raise the return on human capital investment and thereby increase the variance of postschooling investment and log earnings early in the life cycle. This positive effect on relative earnings inequality should decline with experience and possibly become negative. Because of the post–World War II baby boom and baby bust, recent labor force entrants have faced just such a pattern of actual and projected labor force growth rates during their work life. An extended model is tested and generally confirmed using estimates and forecasts of labor force growth rates for the period 1920–2000.

We gratefully acknowledge the comments of James J. Heckman and two anonymous referees, and the assistance of Timothy Maloney and Helen Hamilton Roberts. Financial support was provided by the U.S. Department of Labor, the National Institute of Mental Health, and the Institute for Research on Poverty at the University of Wisconsin—Madison.

[*Journal of Political Economy*, 1984, vol. 92, no. 1]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0005$01.50

I. Introduction

There is substantial evidence of an increase in relative earnings inequality among U.S. males over the last 10–20 years. Henle and Ryscavage (1980) report that the overall and occupation-specific dispersion of annual male earnings in the Current Population Survey (CPS), as measured by the Gini index, has increased steadily over the period 1958–77. An alternative measure of relative inequality from the same data yields a similar result. Plotnick (1982) calculates that the variance of the (natural) log of annual male earnings grew at an annual rate of 2 percent over this period.

Recent changes in the age structure of the labor force provide two simple explanations for this growth in relative earnings inequality. First, both Freeman (1979) and Welch (1979) report a rise in the earnings of experienced workers relative to inexperienced ones and attribute this finding to the labor force entry of the 1950s baby boom. This change would tend to increase the variance of log earnings between experience cohorts. Second, Mincer (1974) has shown that the variance of log earnings within cohorts differs across levels of experience.[1] Hence, the recent "younging" of the labor force may have increased the proportion of adult males in the typically high variance (early) years.[2]

The first objective of this paper is to provide evidence that shows that neither of the straightforward demographic factors above is sufficient to explain the growth in earnings inequality. Data from the CPS indicate nontrivial growth in the variance of log annual (and weekly) earnings within experience cohorts even after the levels of education, experience, and unemployment are controlled for. The second objective is to explain this rise in within-cohort inequality via the response of human capital investment to changes in labor force growth rates. Specifically, it is argued that the children of the baby boom face the prospect of a relatively high rate of return on human capital investment due to the depressed wages caused by their labor force entry and the projected large decline in labor force growth rates during the remainder of this century. A formal model of human capital investment is used to demonstrate how the resulting change in both the average level and variance of human capital investment may account for the recent growth in the dispersion of earnings.

Our primary purpose is to improve our understanding of earnings inequality and, in particular, the influence of labor force growth. In

[1] This result is considered infrequently in discussions of test statistics from studies of wages and earnings, which typically assume a homoscedastic error structure

[2] See Paglin (1975) for a discussion of this issue in the context of the family and personal income distributions.

doing so, however, we extend the scope of the literature on age structure and the earnings distribution by incorporating into our theoretical analysis the impact of demographic phenomena on investment in human capital.

The paper is divided into five parts. In Section II, we briefly review the literature and use the CPS data to demonstrate the increase in relative earnings inequality within labor force cohorts. In Section III, the life-cycle model of human capital investment proposed by Ben-Porath (1967) is used to derive predictions concerning the impact of changes in labor force growth rates on the within-cohort variance of log earnings. In Section IV, these predictions are tested and generally confirmed by the CPS data. Section V provides a summary and conclusion.

II. Review of Previous Analyses and New Estimates of the Life-Cycle Profile of the Variance of Log Earnings

A. *Prior Studies*

Mincer (1974) provides one of the first analyses of the life-cycle behavior of the variance of log earnings among the members of a given labor force cohort.[3] He asserts that the shape of the life-cycle profile will depend on the correlation, within the cohort, between tw vidual characteristics: the log of earnings capacity at the sta market work life (henceforth initial log earnings capacity) a proportion invested of the stock of human capital.[4] Consider composition of the variance of log earnings for a given cohort variance conditional on initial log earnings capacity (the within variance) and the variance between groups of individuals with differ-ent initial log earnings capacities (the between-group variance) Among individuals with identical initial earnings capacities, the observed log earnings profiles of the more intensive investors will exhibit lower intercepts and steeper slopes. The within-group variance declines with experience until "overtaking" and then increases, thus producing a U-shaped life-cycle profile.

---

[3] Chiswick and Mincer (1972) performed a time-series analysis of changes in the *total* variance of log income among U.S. males. Plotnick (1982) has updated this study recently using data on earnings in addition to income. Both studies focus on changes in the variance *between* cohorts defined by education and experience. Specifically, neither study investigates the life-cycle behavior of inequality and, indeed, both explicitly assume that the relative dispersion of earnings within labor force cohorts is independent of the level of experience and education. The validity of this maintained assumption is strongly challenged by the later work of Mincer (1974) and the estimates presented in this paper. Hence, Plotnick's finding of "no significant trend" in the total variance of log earnings may simply reflect the aggregate nature of his measure of inequality combined with the strong secular trends that characterize many of his independent variables.

[4] Throughout this paper, initial earnings capacity will refer to the start of the market work life rather than to the human capital investment life.

The life-cycle profile of the between-group variance depends on the correlation between initial log earnings capacity and the proportion invested of the stock of human capital. The mean log earnings profiles of groups with different initial log earnings capacities will diverge over the life cycle if the proportion invested of the human capital stock is positively correlated with the initial stock. Such divergence will cause the between-group variance to increase with experience. By the same reasoning, a zero (negative) correlation between initial stock and proportion invested would lead to parallel (convergent) profiles and a constant (decreasing) between-group variance with respect to experience.

The life-cycle profile of the variance for the full cohort depends on the pattern and relative sizes of the within- and between-group variances. A sufficiently positive (negative) correlation between the log of initial earnings capacity and the proportion invested of the stock of human capital will result in an increasing (decreasing) total variance over the life cycle. A correlation bounded sufficiently close to zero (see Mincer 1974, p. 103) produces a U-shaped overall profile.

Using annual earnings of white males from the 1/1000 public use sample of the 1960 census, Mincer found that such variance profiles tend to be decreasing in experience for the least educated, U-shaped for high school graduates, and increasing in experience for college graduates. The implication is that the correlation between the log of initial earnings capacity and the proportion invested of human capital is a positive function of education. However, other studies have typically estimated U-shaped profiles with a variety of data sources, for example, Schultz (1975), Smith and Welch (1979), and Hause (1980).

B. Data, Estimation Methods, and Summary Measures

Our first objective is to use data from the CPS for March 1968–79[5] to verify previous findings and to test for the existence of increases in the variance of log earnings within education-experience categories.

Our sample is restricted to civilian males between the ages of 16 and 62, not currently in school (as major activity last week), who either worked 50–52 weeks in the previous year or report the reason for working fewer weeks as something other than being in school or retired. Those self-employed or working without pay are also excluded. Finally, we follow the strategy used by Welch (1979, p. S68) to omit those persons who had, or were likely to have had, earnings levels imputed by the Bureau of the Census.

---

[5] Each survey provides data for earnings in the previous year

Our sample contains 304,177 individual observations from the 12 surveys. For each year, these data were grouped into education-experience cells. The four education categories used were the following: less than 12, 12, 13–15, or 16 or more years of schooling completed. As is well known, the CPS provides no direct information concerning labor force experience. Welch and Gould (1976) used data from the 1940, 1950, 1960, and 1970 censuses of the population and the Coleman-Rossi sample to estimate labor force entry age for males conditional on observed age, education, and birth cohort. These data were kindly provided to us by Finis Welch. Current age minus estimated labor force entry age was used to allocate individuals to (single-year) labor force experience categories.[6] The total number of education-experience cells was 1,960, each one containing at least 50 individual observations.[7]

Table 1 presents the total, the between-cohort, and the average within-cohort variance of log annual and weekly earnings for each year of the sample. All measures exhibit a positive trend, although this is weakest for the within-cohort variance of log weekly earnings. Table 2 provides averages of the within-cohort measures by year and education. The upward trend for weekly earnings is more marked for individuals with less than 16 years of education.

We now turn to regression analysis in order to summar succinctly the relationship between within-cohort earnings and education, experience, and time. It is assumed that the tion variance of log earnings is a linear function of exoge

---

[6] The number of experience cells differed by education level in the following manner less than 12 years of education—46 experience cells, 12 years of education—42 experience cells; 13–15 years of education—39 or 40 experience cells, depending on the survey year; 16 or more years of education—36 or 37 experience cells, depending on the survey year

[7] Sample size precluded separate analyses by race Numerous education-experience cells had less than 10 blacks Smith and Welch (1979), using 1960 and 1970 census data, find that black males have a larger mean variance of log earnings than do whites but similar experience profiles. One solution to this problem would be to use not the observed variance of log earnings as a dependent variable but the variance predicted from the sample regression function of log earnings for each race The latter approach would also permit use of data sources with repeated observations and a larger number of observed determinants of the earnings distribution However, the analysis of the variance of log earnings undoubtedly would become quite sensitive to the functional form of the initial earnings regression Our CPS data provide the advantages of very large sample size and functional flexibility in estimating the conditional variance of log earnings.

TABLE 1

VARIANCE OF ANNUAL AND WEEKLY LOG MALE EARNINGS: 1967–78

	Variance of Log Annual Earnings			Variance of Log Weekly Earnings		
YEAR	Total	Between-Cohort	Within-Cohort*	Total	Between-Cohort	Within-Cohort*
1967	.503	.130	.373	.402	.100	.302
1968	.482	.133	.349	.379	.099	.280
1969	.521	.148	.373	.374	.107	.267
1970	.584	.167	.417	.405	.118	.287
1971	.607	.171	.436	.433	.138	.295
1972	.645	.194	.451	.446	.132	.314
1973	.613	.181	.432	.441	.133	.308
1974	.645	.188	.457	.459	.133	.326
1975	.673	.194	.479	.425	.127	.298
1976	.671	.198	.473	.432	.133	.299
1977	.662	.197	.465	.432	.132	.300
1978	.631	.177	.454	.445	.122	.323

* Weighted average for cohorts defined by four educational levels and by single years of labor force experience

ables. Hence, the sample log variance can be expressed in the following forms:

$$\ln \operatorname{var} AE_{it} = \alpha_0 + \alpha_1 \exp_{it} + \alpha_2 \exp_{it}^2 + \alpha_3 U_t + \alpha_4 t + \alpha_5 t \times \exp_{it} + \epsilon_{1it}, \tag{1}$$

$$\ln \operatorname{var} WE_{it} = \beta_0 + \beta_1 \exp_{it} + \beta_2 \exp_{it}^2 + \beta_3 U_t + \beta_4 t + \beta_5 t \times \exp_{it} + \beta_6 PT_{it} + \epsilon_{2it}, \tag{2}$$

where *AE* and *WE* are annual and weekly real earnings for the *i*th experience cell in the *t*th year, *U* is the male unemployment rate, *t* is a time trend, and *PT* is the proportion of normally part-time (week) workers. If earnings are lognormally distributed, then ϵ_{1it} and ϵ_{2it} have a gamma distribution in small samples but an (approximate) asymptotic normal distribution.[8]

---

[8] The sample variance is constrained to be nonnegative. Hence the asymptotic distribution can never be exactly normal, but the central limit theorem justifies this approximation. The variance of log earnings obviously can only be estimated for nonzero earners. One characteristic of our sample is the growth over time in the proportion of otherwise valid observations with zero earnings. For the higher educational categories, this proportion is relatively small. For high school dropouts, however, the percentage of zero earners rose from 6 percent in 1968 to 16 percent in 1979. As shown by Heckman (1979), the mean and variance of a censored sample (nonzero earners) generally provide inconsistent estimates for the corresponding moments of the *potential* earnings distribution among all individuals. Heckman also provides an estimation strategy to correct for such sample selection bias assuming a homoscedastic error structure in the full population. However, Hurd (1979) has shown not only that heteroscedastic

TABLE 2

AVERAGE WITHIN-COHORT VARIANCE OF ANNUAL AND WEEKLY LOG MALE EARNINGS BY YEARS OF SCHOOLING: 1967–78

	VARIANCE OF LOG ANNUAL EARNINGS				VARIANCE OF LOG WEEKLY EARNINGS			
YEAR	11 Years or Less	12 Years	13–15 Years	16 Years or More	11 Years or Less	12 Years	13–15 Years	16 Years or More
1967	.511	.272	.259	.314	.390	.229	.221	.284
1968	.483	.250	.275	.288	.365	.212	.233	.254
1969	.539	.283	.260	.269	.356	.204	.210	.245
1970	.589	.320	.290	.366	.376	.212	.221	.317
1971	.589	.340	.366	.389	.385	.221	.261	.302
1972	.631	.368	.396	.342	.416	.252	.279	.289
1973	.620	.359	.355	.327	.421	.257	.267	.253
1974	.664	.385	.405	.332	.434	.279	.308	.274
1975	.690	.410	.410	.378	.412	.233	.283	.282
1976	.672	.424	.404	.361	.411	.257	.261	.262
1977	.696	.400	.400	.355	.405	.257	.275	.271
1978	.663	.420	.356	.370	.430	.283	.278	.320

NOTE.—Weighted average for cohorts defined by four educational levels and single years of labor force experience.

The weekly earnings equation, when the proportion of part-time (week) workers is controlled for, provides the best available method of estimating a model for the variance of the log of hourly wages with CPS data.[9]

C. *Initial Regression Estimates with CPS Data*

Table 3 presents the weighted regression estimates of equations (1) and (2). The first set of estimates were obtained with all observations using dummy variables for educational categories (D12, D13–15, D16+) Eleven years of schooling or less is the omitted category. The subsequent estimates permit a full range of education interactions.[10]

Unlike Mincer (1974), we find strong evidence of a U-shaped experience profile for all levels of education. The only exception is weekly earnings among college graduates. Hence, as with Schultz (1975), Smith and Welch (1979), and Hause (1980), our results indicate a weak correlation between initial earnings capacity and the proportion invested of the stock of human capital. The less-educated groups exhibit larger intercepts and more sharply curved experience profiles This indicates a greater diversity of initial earnings capacity and/or investments. An additional factor suggested by Mincer is more extensive job search among those with fewer years of schooling.

Increases in the unemployment rate raise the variance of log annual earnings. A likely reason for this appears to be an increase in the variance of weeks worked during recessions, since unemployment has a negative, though generally weak, effect on the variance of log weekly earnings.

---

errors result in inconsistent coefficient estimates, but that moderate variation in the error variance may lead to substantial asymptotic bias. Our model explicitly focuses on the structure of heteroscedasticity and, hence, an alternative estimation procedure is required Dooley (in press) shows that, if one assumes a bivariate lognormal distribution of offered and reservation earnings conditional on experience and education, estimates of the two uncensored moments of the offered earnings distribution can be derived from the sample proportion of zero earners and the first three censored sample moments This method was used to estimate the uncensored variance of the log of annual and weekly earnings for each education-experience cell. However, there were no significant qualitative differences between the regression results obtained with the censored (observed) and the uncensored variance estimates. Hence, we present only the former in the text.

[9] The overall proportion of part-time workers never exceeds 4 percent, and this proportion is trivial among those with at least 13 years of education. Hence we believe that the robust nature of our results across educational levels lends support to the empirical strategy we have adopted.

[10] It should be noted that the proportion of observations excluded because of imputed earnings did rise substantially over the observation period; however, the inclusion of this variable in eqq. (1) and (2) did little to alter the estimates in table 1 This variable's own coefficients tend to be weakly positive in the case of annual earnings and nonsignificant for weekly earnings

The trend coefficients are of particular interest. These linear effects are significantly positive for the entire sample and for each educational level except the college graduates. The consistently negative interactive effects between trend and experience, again with the exception of college graduates, indicate that the growth in inequality is greater among the young. However, the test statistics for these estimates usually fall short of standard threshold levels. The total trend effect remains positive up to, or beyond, the normal age of retirement. When measured as a semi-elasticity, that is, the annual percentage increase in the variance of log earnings, the magnitudes are not trivial. The semi-elasticities range from 2 percent to 4 percent per year for all but the most educated group.[11]

Hence, the most important implication to be derived from table 3 is that the increase in earnings inequality among all males reported in Section I is not just a result of changes in the relative earnings of young to old workers or of growth in the proportion of young (high variance) workers. Indeed, the 2 percent average annual rate of increase in the total variance of log earnings frequently *understates* growth for the within-cohort variance conditional on human capital and cyclical factors, particularly for younger workers.

Several possible explanations for these trends appear not to be fully compatible with the pattern of results in table 3. First, the st positive trends for weekly earnings among those with 12–15 education imply that this phenomenon is not merely a labo response on the part of low-wage individuals to the in generosity of government transfer programs over this perio the significant trend coefficients limited to annual earnings high school dropouts, the explanation above might be appeali such is not the case.

A second possibility is an increase in the variance of the initial stock of human capital of successive cohorts *within* our educational categories. One source of such a change would be an increase in the variance of years of schooling. However, the middle groups offer little or no opportunity for such changes, and we suspect that the dispersion in years of schooling among high school dropouts has fallen. Plotnick (1982) shows that the total variance of years of schooling decreased

---

[11] When the variance of the log of weeks worked is included in the regression for annual earnings, the resulting coefficient is highly significant, with a coefficient approximately equal to two. The time trends drop by roughly 20 percent but retain significance. The effect of male unemployment becomes significantly negative for the two lowest education groups. The major inferences drawn throughout the paper are not altered. We have not reported these results for two reasons: our weekly earnings regressions control for variation in weeks worked in the most satisfactory fashion, and we consider the issue of employment stability as one of the returns to human capital investment to be an open question.

TABLE 3

Regression Estimates of Determinants of Within-Cohort Variance of Annual and Weekly Log Male Earnings, 1967–78

	Intercept and Schooling Dummies	Exper	Exper^2	Male Unemployment	Trend	Trend × Exper	Part-Time	R^2
Annual earnings.	.69	−.024	.00052	.016	.010	−.000087	...	.40
	(29.6)	(15.2)	(16.8)	(3.4)	(3.6)	(0.9)		
D12	−.26	...	...	...	...	...	...	...
	(24.9)							
D13–15	−.27	...	...	...	...	...	...	...
	(19.7)							
D16+	−.28	...	...	...	...	...	...	...
	(20.7)							
Weekly earnings.	.34	−.0022	.000081	−.006	.006	−.000031	.80	.32
	(18.1)	(1.7)	(3.2)	(1.8)	(2.9)	(.4)	(14.9)	
D12	−.13	...	...	...	...	...	...	...
	(16.1)							
D13–15	−.12	...	...	...	...	...	...	...
	(16.1)							
D16+	−.09	...	...	...	...	...	...	...
	(8.6)							

Education 11 years or less								
Annual earnings	87	−.041	00080	018	.021	−.00045	. . .	.44
	(18.8)	(14 6)	(15.5)	(1.9)	(3 7)	(2 6)		
Weekly earnings	46	− 011	.00023	− 003	0070	− 00013	.62	.33
	(13.6)	(5.2)	(5.6)	(5)	(1 8)	(1 1)	(8.0)	
Education: 12 years.								
Annual earnings	.41	− 026	.00064	012	015	−.00029	. . .	.37
	(13.5)	(11.6)	(13.2)	(1 8)	(4.1)	(2.0)		
Weekly earnings	.25	−.007	00021	− 011	0094	−.00010	.65	.28
	(11.1)	(4.0)	(5 6)	(2.5)	(3 8)	(1.0)	(7 0)	
Education: 13–15 years:								
Annual earnings	37	−.024	00065	.020	.0093	− 00024		.18
	(8.4)	(6.9)	(8 1)	(2.3)	(1 8)	(1.0)		
Weekly earnings	18	−.003	00017	− 0011	0080	−.00016	.59	17
	(4.8)	(1.4)	(2 3)	(2)	(2.1)	(.9)	(5.8)	
Education. 16 years or more:								
Annual earnings	.37	−.023	00078	017	0004	.0001	. . .	.12
	(6 6)	(4.7)	(6 4)	(1.5)	(.1)	(.3)		
Weekly earnings	16	.0022	00017	− 0013	− 00028	0001	1.5	14
	(2 6)	(4)	(1 2)	(.1)	(0)	(2)	(4.3)	

NOTE —Asymptotic normal statistics in parentheses over the period 1958–77. Increases in the variance of schooling quality or innate "ability" are additional sources of growing dispersion in initial stocks. The CPS unfortunately offers no information relevant to this question. However, the frequently heard claim that passing standards have fallen at all levels of education generally would serve to inflate the variance of ability most strongly in our highest educational category, the college graduates. Yet it is for this group alone that *no significant* trend is found.

A third potential source of change in within-cohort inequality is the higher tax rates imposed by recent inflationary increases in nominal earnings. The effect of such "tax creep," however, is not clear a priori. Assume, for example, an adjustment of the pretax real wage structure sufficient to maintain constant relative levels of posttax real wages due to constant relative costs of training. Further assume that effective marginal tax rates increase at a decreasing rate with nominal earnings. An increase of x percent in all nominal earnings may lead to increases in marginal and average tax rates that are a negative function of initial earnings. Individuals with the highest initial earnings receive the largest nominal raises but experience the smallest tax rate increases per dollar. Hence, the posttax relative wage structure may expand and, under the adjustment process assumed above, induce a compression in the pretax relative wage structure. Only under the assumption of a convex effective marginal tax rate schedule can a strong case be made for a positive effect of inflation on the relative inequality of real wages, but there is little support for this assumption.

III. The Effect of Changes in Labor Force Growth Rates

Mincer's (1974) interpretation of the life-cycle profile of the variance of log earnings centers on the distribution of investment in human capital within a given cohort. Hence, a logical starting point for an investigation of intercohort differences in the variance of log earnings is changes in the distribution of investment across cohorts. We focus on changes in labor force growth rates as a source of such changes in investment and earnings. In essence, we assert that the baby boom cohorts have faced relatively high rates of return to human capital investment, particularly on-the-job training, due to the depressed wages caused by their market entry, abetted by the prospect of rapid deceleration in labor force growth during the remainder of this century. Such increases in investment incentives are shown to increase within-cohort earnings inequality early in the life cycle.

Our approach is similar to that taken by Freeman (1979) and Welch (1979) in that we use demographic phenomena to explain recent modifications in the earnings distribution. A major distinction is that we consider voluntary human capital investment as an important avenue of such modifications. Previous studies, in contrast, concentrate on concurrent changes in the relative supplies and wages of young versus old workers.

Our theoretical results are consistent with those derived and empirically confirmed by Welch in particular. However, the current approach permits use of a formal model of human capital accumulation, a feature we believe useful for a satisfactory analysis of earnings dispersion.

In this section, we first utilize a standard model of optimal human capital accumulation provided by Ben-Porath (1967) to demonstrate that inequality of observed log earnings within a cohort is a function of expected growth in the rental rate of human capital. Next, results from standard neoclassical growth models are used to establish the relationship among growth rates in the labor force and the rental rate for human capital.

A. *Investment in Human Capital and the Variance of Log Earnings*

Ben-Porath (1967) provides a well-known model of the production of human capital and the life cycle of earnings. Decision makers assumed to be concerned solely with wealth, that is, the pr
of disposable earnings. Furthermore, human capital is ass
homogeneous in contrast to the models of Freeman and W
assumptions of wealth maximization and homogeneity, whi
satisfactory, do provide a theoretical framework for investm
sis that is both tractable and conducive to plausible predictio
14 for a brief discussion of results from a model of utility m
imization.)

Human capital is produced according to the following function

$$Q_t = \beta_0(S_tK_t)^{\beta_1}, \quad 0 \leq S_t \leq 1, \quad 0 \leq \beta_1 \leq 1, \tag{3}$$

where Q_t is the flow of human capital produced; K_t is the current stock of human capital; and S_t is the fraction of K_t devoted to the production of human capital, so that S_tK_t is the quantity of human capital utilized for investment.

The human capital production function originally used by Ben-Porath included purchased inputs. Our focus is on postschooling investment ($S_t < 1$) and we therefore adopt the standard assumption that own time is the sole input in (3). For convenience, we assume also zero depreciation of human capital.

Ben-Porath shows that the optimal production level of human capital production during the postschooling phase can be expressed as follows:

$$Q_t = (\beta_0)^{1/(1-\beta_1)} \left(\frac{\beta_1}{\psi}\right)^{\beta_1/(1-\beta_1)} [1 - e^{-\psi(T-t)}]^{\beta_1/(1-\beta_1)}, \tag{4}$$

where $\psi = r - \eta$; r = interest rate; η = exponential growth rate of the human capital rental rate, which has the initial value α_0; and T = exogenously determined terminal year. Observed earnings, $\hat{E}_t$, can be expressed as the difference between earnings capacity, E_t, and forgone earnings, I_t. Hence

$$\hat{E}_t = E_t - I_t, \tag{5}$$

where $E_t = \alpha_t K_t$, $I_t = \alpha_t S_t K_t$, and $\alpha_t = \alpha_0 e^{\eta t}$.

In order to generate a distribution of observed earnings within a labor force cohort, we follow Mincer in assuming that individuals can be distinguished from one another on the basis of two characteristics: initial log earnings capacity, ln E_0 ($= \alpha_0 K_0$), and the ability to produce human capital (or learning ability), which we shall assume to be represented by β_0. If all individuals face an equal-length work life, then, for any given ln E_0, higher levels of β_0 lead to greater investment. Hence, conditional on ln E_0, β_0 is negatively (positively) correlated with the intercepts (slopes) of observed log earnings profiles.

We now proceed to show that the variance of log earnings is a function of η, the growth rate of the human capital rental rate, but not of α_0, the initial value of the rental rate. The first step is to demonstrate the effect of changes in η and α_0 on observed log earnings. It is tedious, but straightforward, to show that

$$\frac{\partial \ln \hat{E}_t}{\partial \alpha_0} = \frac{1}{\alpha_0} \tag{6}$$

and

$$\frac{\partial \ln \hat{E}_t}{\partial \eta} = t + \frac{\delta_t}{(1 - \beta_1)\hat{E}_t}, \tag{7}$$

where $\delta_t = \alpha_t \beta_1 \int_0^t C_t Q_t - C_t I_t$ and

$$C_t = \frac{1}{\psi} - \frac{(T - t)}{e^{\psi(T-t)} - 1} > 0.$$

Equation (6) shows that the effect on log earnings of a change in α_0, the level of the rental rate profile, is a function only of α_0. All log earnings profiles shift up or down by the same amount. Relative levels of earnings, and hence the variance of log earnings, are unaffected

Equation (7) shows that more rapid growth in the rental rate lower ln $\hat{E}_t$ early in the life cycle and increases ln $\hat{E}_t$ later on; that is, a higher rate of return stimulates greater investment. To see this, note that δ_t is negative at $t = 0$ when $Q_t = K_t - K_0 = 0$ and positive at $t = T$ when $I_t = 0$.[12]

Next we take the partial derivatives of (7) with respect to β_0 and ln E_0:

$$\frac{\partial^2 \ln \hat{E}_t}{\partial\eta\partial\beta_0} = \frac{\delta_t\alpha_t K_0}{(1 - \beta_1)^2\beta_0\hat{E}_t^2} \tag{8}$$

and

$$\frac{\partial^2 \ln \hat{E}_t}{\partial\eta\partial \ln E_0} = \frac{-\alpha_t K_0\delta_t}{(1 - \beta_1)\hat{E}_t^2}. \tag{9}$$

Equations (8) and (9) each switch signs at the point where $\delta_t = 0$. The value of (8) is negative early in the life cycle and positive later on. The opposite is true of (9).

The results in (7), (8), and (9) are sufficient to establish the qualitative effect of a change in the rental rate growth rate on the variance of log earnings. A formal analysis is available on request from the authors. In the following discussion, we present our basic results using a more intuitive, graphical approach.

As in Section II*A*, it is convenient to decompose the variance of log earnings, and changes therein, into the variance conditional on ln F, (the within-group variance) and the variance across groups
ferent values of ln E_0 (the between-group variance). Assu
that ln E_0 (K_0) and β_0 are distributed independently.

Figure 1 depicts changes in the within-group variance
solid lines represent the log earnings profiles at a low value
individuals with a common initial stock of human capital
different levels of β_0. The more able learner (ln $\hat{E}_t : \beta_{0H}$) inv
intensively, exhibiting lower initial earnings and more rapid grow

The broken lines represent the optimal profiles of the same individuals given an increase in η. Equation (9) shows that both profiles rotate counterclockwise, reflecting more intensive investment, with the greatest response elicited from the more able learner (ln $\hat{E}_t' : \beta_{0H}$). The rise in η has increased the within-group variance of investment.

---

[12] Equation (7) is equal to zero where

$$\frac{C_t I_t}{(1 - \beta_1)\hat{E}_t} = t + \frac{1}{(1 - \beta_1)\hat{E}_t}\left(\alpha_t\beta_1\int_0^t C_t Q_t\right).$$

It can be shown that $\partial C_t/\partial t < 0$ for all t. Note that eq. (7) may stay negative for some time, since C_t can be quite large at low t. For example, for $\psi = .02$, $C_t = 19.2$ at $T - t = 45$ and $C_t = 13.5$ at $T - t = 30$.

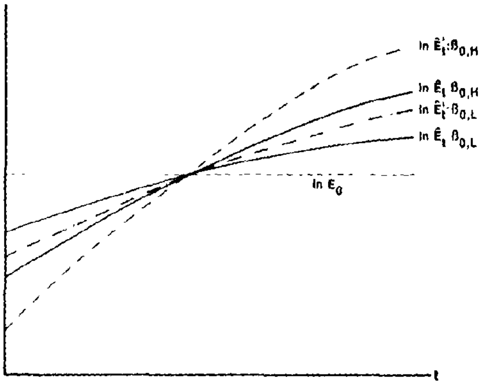


Fig. 1.—Human capital rental rate growth and the log earnings profile for given initial log earnings capacity, where β_0 = "learning ability"; η = growth rate of human capital rental rate, $\ln E_0$ = initial log earnings capacity; $\ln \hat{E}_t \cdot \beta_{0,H}$ = log earnings given high β_0 and low η; $\ln \hat{E}_t^{\prime} . \beta_{0,H}$ = log earnings given high β_0 and high η; $\ln \hat{E}_t : \beta_{0,L}$ = log earnings given low β_0 and low η; $\ln \hat{E}_t^{\prime} : \beta_{0,L}$ = log earnings given low β_0 and high η

Hence, the within-group variance of log earnings grows at all points with the possible exception of a short interval in the middle years. Variance growth is most pronounced for very low and very high t.

Figure 2 depicts the change in the between-group variance. The concave solid lines depict the mean log earnings profiles, at a low value of η, for two groups of individuals with identical distributions of β_0 but different values of $\ln E_0$. Note that the profile for the group with lower initial earnings capacity, $E(\ln \hat{E}_t | \ln E_{0,L})$, is more steeply sloped than $E(\ln \hat{E}_t | \ln E_{0,H})$. This follows from the "neutrality" assumption that human capital production (Q_t) is independent of the initial earnings capacity combined with the assumption of an independent distribution of $\ln E_0$ and β_0. Individuals with lower values of $\ln E_0$ on average invest a greater proportion of their earnings capacity and hence exhibit more rapid growth in $\ln \hat{E}_t$.[13]

The dashed lines again provide profiles corresponding to an increase in η, the growth rate of rental rates. As indicated by equation (9), each log earnings profile rotates counterclockwise, reflecting greater average investment in human capital. The average increase in investment is independent of initial earnings capacity, just as were the initial levels of investment. Hence, the between-group variance of

---

[13] The Ben-Porath model implies, then, that with $\ln E_0$ and β_0 distributed independently, the variance of log earnings between groups with different $\ln E_0$ should decrease over the life cycle. This prediction is not inconsistent with the U-shaped experience profiles in table 3, especially since the troughs occur at a relatively high level c experience, generally 15–25 years.

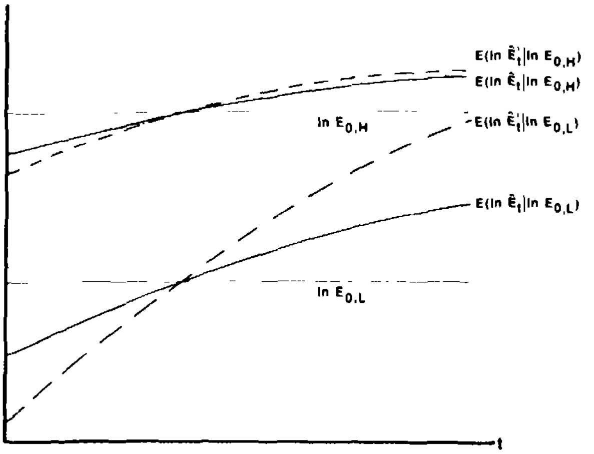


FIG. 2.—Human capital rental rate growth and the expected log earnings profile for different initial log earnings capacities, where η = growth rate of human capital rental rate, $\ln E_{0,H}$ = high initial log earnings capacity, $\ln E_{0,L}$ = low initial log earnings capacity, $E(\ln \hat{E}_t|\ln E_{0,H})$ = expected value of log earnings given $\ln E_{0,H}$ and low η, $E(\ln \hat{E}'_t|\ln E_{0,H})$ = expected value of log earnings given $\ln E_{0,H}$ and high η, $E(\ln \hat{E}_t|\ln E_{0,L})$ = expected value of log earnings given $\ln E_{0,L}$ and low η, $E(\ln \hat{E}'_t|\ln E_{0,L})$ = expected value of log earnings given $\ln E_{0,L}$ and high η.

investment does not change, but the between-group varia
earnings does change. The *proportional* decrease in mea
(the change in $E[\ln \hat{E}_t|\ln E_0]$) early in the life cycle is greate
with lower initial earnings capacity. The same holds true t
tional increases in mean earnings in the later years. The
$E(\ln \hat{E}_t|\ln E_{0,L})$ at high t is larger than that in $E(\ln \hat{E}_t|\ln$
consequence, the between-group variance of log earnings
early in the life cycle. However, this positive effect diminishes w
experience and eventually becomes negative.

In summary, more rapid growth in human capital rental rates, by raising the mean and variance of human capital investment, leads to increases in both the within- and between-group variance of log earnings early in the life cycle. At later ages, the effects are offsetting, within-group variance tends to grow and between-group variance tends to shrink.[11]

---

[11] Two comments on extensions of our model are relevant. First, it can be shown that the partial equilibrium effects of a higher proportional income tax rate are identical to those of a more rapidly growing rental rate in a wealth maximization model. Second, we have explored the effect of changes in the growth rate of the human capital rental rate within the context of Heckman's (1976) life-cycle model of utility maximization. In such a model, earnings profiles are affected by both human capital investment and leisure consumption. All of our results for human capital investment continue to hold in the Heckman model. A more rapidly growing rental rate raises the value of leisure

The assumption that $\ln E_0$ (or K_0) and β_0 are independently distributed has simplified substantially the preceding analysis. Our interest for empirical purposes is in the joint distribution of β_0 and $\ln E_0$ conditional on a given level of schooling. It can be shown that the assumption of an independent distribution is not inconsistent with the Ben-Porath model. (A formal analysis is available on request from the authors.) In addition, the consistently U-shaped experience profiles in table 3 lend empirical support to this assumption, as do most previous studies.

The basic implications of alternative assumptions may be outlined briefly. Our analysis of the within-group variance is conditional on $\ln E_0$ and therefore unaffected. The assumption that $\text{cov}(\beta_0, K_0) < 0$ will tend to reinforce the predictions above for the effect of η on the between-group variance. If $\text{cov}(\beta_0, K_0) > 0$, then our predictions for the between-group variance may not necessarily hold. It should be emphasized, however, that there is little empirical support for this last assumption. We assume, then, that $\text{cov}(\beta_0, K_0)$ is so small in algebraic value that the effect of a rise in the growth rate of the human capital rental rate on the variance of log earnings is positive in the early years but declines and possibly becomes negative at higher levels of experience.

B. The Role of Labor Force Growth

The previous section analyzed the relationship between rental rate growth and the variance of human capital investment and log earnings. The predictions derived are quite difficult to test, however, given the (directly) unobservable nature of rental rates. The purpose of this section is to establish a link between an observable variable—labor force growth rates—and rental rate growth. Our basic assertion is that changes in the growth rate of the labor force should be negatively associated with the growth rate of rental rates and, hence, the return on human capital investment. Given the analysis of Section III*A*, this proposition establishes a relationship between observables,

---

consumed (depresses observed earnings) at early ages in the case of a Cobb-Douglas or CES utility function. At later ages, the effect cannot be signed. The effect at early ages is stronger for the more able learners (higher β_0), thereby serving to increase the within-group variance of log earnings. The effect at early ages is also stronger for those with higher initial earnings capacity. This serves to decrease the between-group variance of earnings, however, the effect for log earnings cannot be signed. In summary, the incorporation of leisure consumption tends to reinforce the conclusions in the text where the comparative dynamics can be signed, but such an exercise frequently yields unsigned results.

that is, estimates and projections of labor force growth and the variance of log earnings. The empirical relevance of this relationship is discussed in Section IV where it is shown that the post–World War II baby boom and baby bust can be expected to induce an unprecedentedly high rate of deceleration in labor force growth over the last quarter of this century.

Consider first the basic one-sector neoclassical model of economic growth (Solow 1970; Burmeister 1980). A well-known result of this model is that the steady-state level of real wages is negatively related to the rate of labor force growth. In the absence of technological change, the real wage is constant in the steady state. A decrease in the rate of labor force growth leads to a new steady state with a higher capital to labor ratio and real wage. Hence, a drop in the labor force growth rate induces a temporary increase in the rate of wage growth from zero to some positive number.

The incorporation of labor-augmenting technological change leads to a similar result for the wage rate measured in efficiency units of labor. In the steady state, wages per efficiency unit (rental rates) are constant and wages per worker grow at a rate equal to the rate of technological change. A decline in the labor force growth rate leads to a higher level wage per efficiency unit and a growth path for wages per worker with a higher intercept but unchanged slope. Therefore, a permanent drop in the labor force growth rate brings about a [illegible]rary increase in the growth rate of wages per efficiency un[illegible] worker. Changes in the growth rate of the labor force and [illegible] negatively related.

The assumptions concerning individual worker behav[illegible] one-sector growth model are far different from those in [illegible] Porath model. In the former, workers make no decisions. E[illegible]nously given (and costless) growth rates in labor supply and technology determine the wage per efficiency unit (the rental rate) and per worker. In the Ben-Porath model, the level and growth of the rental rate are given and determine the individual's optimum paths for wages and supply of human capital to the market.

A more appropriate theoretical framework for our analysis would be a multisector growth model that incorporates the individual wealth-maximizing decisions of successive cohorts of workers. Anticipated labor force and technological growth along with other exogenous factors would determine the distribution of human capital investment and, ultimately, the equilibrium growth paths for the rental rate, the observed wage distribution, and the aggregate supply of human capital. To our knowledge, such a general equilibrium model of overlapping generations of workers remains on the profession's agenda for future research. To accomplish this task is beyond the scope of our present efforts.[15]

We suggest, however, that the implications of the one-sector growth and Ben-Porath models are both intuitively appealing and indicative of the comparative dynamic results one would obtain from a more extensive theoretical treatment. In such an expanded model, the critical relationship, for our purposes, would be that between labor force growth rates and the rate of return on human capital investment. Declining rates of labor force growth should lead to deceleration in the growth of the aggregate stock of human capital and acceleration in rental rate growth—for any given distribution of human capital investment. The consequent increase in the expected rate of return for any given level of training would stimulate investment among labor force entrants and, given the assumptions of the Ben-Porath model, increase the within-cohort variance of log earnings during the initial years of experience, with a diminishing and possibly negative effect at later stages.[16] The general equilibrium effects would, of course, be more complicated. However, we would not anticipate a change in the qualitative relationship among labor force growth rates, investment, and the variance of log earnings.

IV. Estimates of the Impact of Changes in Labor Force Growth Rates

In this section, we extend our empirical analysis to include the impact of past and projected changes in the rate of labor force growth. We show that post–World War II fertility trends can be expected to result in large drops in labor force growth rates over the next 20 years. This phenomenon, combined with the preceding analysis, provides a possible explanation for the recent growth in earnings inequality.

---

[15] Razin (1972) offers one of the few models of economic growth with endogenous investment in human capital. For several reasons, however, this model is inappropriate for our purposes. Most important, the decision-making units are infinitely lived households that grow in size at a fixed rate and *costlessly* pass on acquired earnings capacity to new members. The individual analogue is that of a person with infinite life for whom the number of available hours per period grows at a fixed rate. The role of "population growth" in Razin's framework is quite different from one that we believe to be appropriate for analysis of such phenomena as the baby boom and baby bust.

[16] We believe that the assumption of perfectly homogeneous human capital, though convenient and standard in systematic investment models, is not necessary to arrive at our predictions of relatively high rates of return among cohorts anticipating future deceleration in labor force growth. If the elasticity of substitution among workers of different experience levels is zero, then our assertions are unwarranted, but this assumption is at least as unrealistic as that of an infinite elasticity. What is necessary is the assumption that this elasticity of substitution is sufficiently high for there to be negative correlation between the cohort's expected rate of return on investment and the anticipated *changes* in labor force growth rates.

TABLE 4

AVERAGE ANNUAL RATES OF GROWTH IN CIVILIAN LABOR FORCE, AGES 16-64

		By Years of Schooling			
Years	Total	11 or Less	12	13–15	16 or More
1920–30	1.63	1.18	3.12	3.11	3.21
1930–40	1.44	.48	4.09	3.87	3.16
1940–50	1.10	.24	3.35	3.34	3.71
1950–60	1.16	−1.03	3.77	3.12	4.37
1960–70	1.79	−1.73	3.94	5.07	4.14
1970–80	2.46	−2.72	2.94	5.75	6.44
1980–90	1.60	−3.76	.80	3.21	5.15
1990–2000	.96	−6.95	−.81	2.15	4.05

SOURCE see App

Table 4 presents 10-year averages of the estimated and projected annual rates of change in the civilian labor force, ages 16–64, from 1920 to 2000. The data and methods used to derive these figures are described in detail in the Appendix. The first column presents the averages for the entire labor force. The decline in these growth rates from 1920 through the mid-1950s reflects primarily decreases in the rate of adult population growth due to curtailed immigration and fertility. The next 20 years witnessed an increase of almost centage points due primarily to the maturation of the boom aided by rising rates of female labor force deceleration of similar magnitude (1.5 points) is projected 2 decades of this century, primarily because of the baby the rapid decline of fertility rates during the 1960s and

The data in table 4 indicate a substantial variation in these education level. The projected drop in growth rates is common groups. The acceleration in labor force growth during the 1960s and 1970s, though, was an increasing function of years of schooling, with the number of high school dropouts continuing to shrink. However, we believe that heavy emphasis on the educational differences in table 4, particularly among the three lowest categories, is *not* warranted, since we lack strong a priori knowledge concerning elasticities of substitution among schooling groups and intertemporal changes in schooling content.

Our central point is best conveyed by the data for the total labor force. Over the observation period, 1967–78, our CPS sample was increasingly populated with individuals who face the prospect of sharply decreasing rates of labor force growth during the first 20–30 years of their work life. The children of the baby boom entered the labor force during a period of rapid growth when rental rates were depressed and, hence, investment cheap. The returns on such investment will be reaped during a period of slow labor force growth and correspondingly higher expected rental rates. The resulting stimulus to rates of return on postschooling investment would be predicted by our model to increase the variance of log earnings among young workers during our sample period. Therein lies one possible explanation for the increases in the within-cohort variance of log earnings exhibited in table 3.

An additional factor may be those individuals in our sample who entered the labor force in the late 1940s and 1950s. Such persons entered the labor force during a period of slow growth and faced the prospect of accelerating rates of labor force growth during the first 15–20 years of experience. Modest growth in earnings inequality during our sample period for those cohorts in the middle stages of their work life is also consistent with our model and the results presented in table 3.

It is interesting to note that the implications of our model are not inconsistent with those of Welch (1979), even though the latter focuses on the effect of labor force age composition on the relative wages of young to old in a model with heterogeneous human capital and no voluntary investment.[17] Welch asserts that the wage-depressing impact of large cohort size is felt most strongly in the early years of the work life and diminishes with experience. Hence, the recent large cohorts of labor force entrants should exhibit, and in Welch's findings do exhibit, earnings profiles with lower intercepts and steeper slopes than those of smaller cohorts. Our model predicts the same; however, the underlying reasons are temporarily depressed rental rates *and* increased postschooling investment.[18]

[17] Clearly, a more satisfactory theoretical analysis would combine the assumptions of heterogeneous human capital and endogenous investment However, the absence from the literature of such models testifies to the substantial obstacles represented by mathematical intractability and the absence of a priori guides as to the appropriate skill categories

[18] Additional corroborating evidence on increased postschooling investment among recent labor force entrants is scarce Stafford and Duncan (1980) report that the proportion of hours on the job that were devoted to work declined substantially for males between 1965 and 1975 Unfortunately, the authors did not have sufficient data to determine if the released hours were devoted to training or leisure A potential problem with our analysis is that, within the context of the Ben-Porath model, an increase in the rate of return on investment should also prolong the period of specialization, i e., increase levels of formal schooling. However, Freeman (1981) has shown that the proportion of civilian male high school graduates entering college dropped over our sample period The proportion of civilian males, aged 18–19, enrolled in college shows a similar decline. Of course, this drop in enrollment rates may reflect a shift from college to other forms of training rather than a reduction in the overall level of investment. There are several arguments to support the former interpretation, i e , that of a change in the *form* of training First, most of the drop in civilian male enrollment rates occurred between 1969 and 1974 During this period, the proportion of males in

A more systematic test of our model requires extension of the regression analysis in Section II. For this, it was necessary to construct a measure of anticipated changes in labor force growth rates for each cohort in our sample. The measure selected was a straightforward one—the simple time trend for annual rates of growth of the civilian labor force ages 16–64 (henceforth LFGRT) during the first 20 years of each cohort's work life. For example, in the case of cohorts entering the labor force in 1950, LFGRT is the slope from a regression on time of the annual rates of change in labor force size during the period 1950–70 (see the App. for further details). Our prediction is that LFGRT should have a negative linear effect on the variance of log earnings and a positive interaction with experience.

Our formal theoretical model employed the simplifying assumption of homogeneous human capital. This implies that LFGRT should be estimated using data for the entire labor force. Female participants were included both in the spirit of this assumption and because there is a strong correlation between estimates of LFGRT based on the total and male labor force.[19] Table 4 does indicate variation in labor force growth rates by level of education. For reasons stated above, we believe that it is misleading to attach heavy emphasis to these differences. However, as a sensitivity test, the regressions discussed below were estimated with measures of LFGRT calculated both by education category and for the labor force as a whole.[20]

For most schooling groups and the labor force as a whole
tends to have negative values for the oldest and youngest co
positive values for workers in the middle portion of their w
LFGRT is negative for all cohorts with less than 12 years of
and greatest in absolute value for the most recent ent
length of the projection period for LFGRT is obviously a
judgment. The Bureau of Labor Statistics currently provided t

---

military service declined substantially. Second, Freeman considers only college enrollment as opposed to all types of postsecondary formal training. A consideration of enrollment measures that incorporate these two factors paints a different picture. Specifically, the proportion of total males ages 18–24 enrolled in postsecondary education was virtually unchanged between 1969 and 1974, whereas the decline in the enrollment rate for civilian males was 5 percent (Zsigmond et al. 1978). Finally, Freeman shows that the direct real cost of higher education (tuition and fees) rose between 1969 and 1974 and was constant between 1969 and 1978. This fact, combined with falling wages (rental rates), would serve to lower the cost of the more own-time intensive on-the-job training (OJT) relative to formal schooling.

[19] Note, however, that we still seek to explain the variance of log earnings for males. Even less research has been devoted to female earnings inequality than to male. Among the factors that complicate this type of analysis for women are their low participation rates (relative to men) at a point in time, the importance of part-time work, and the frequently intermittent nature of market work over the life cycle.

[20] The measures of within-cohort earnings inequality are still measured conditional on years of schooling, as called for by our theoretical analysis in Sec. IIIA.

casts to the year 2000. The use of 10- and 15-year trends led to few qualitative changes in the regression results reported below.[21]

The use of LFGRT as a measure of expected growth in human capital rental rates is clearly open to criticism. We have assumed that individuals are forward looking. In order to support our predictions, however, one need not assume perfect forecasts of labor force growth by market entrants during our sample period. A sufficient condition is the recognition on the part of such entrants that they faced poor initial earnings prospects, relative to their predecessors, combined with the expectation of substantial improvement in labor market conditions in the 1980s and 1990s.[22]

Table 5 presents the regression estimates obtained when LFGRT and LFGRT × experience are added to equations (1) and (2). (The education-specific results are those obtained with the education-specific measure of LFGRT.) The coefficients for these variables are invariably in accord with theoretical predictions and generally have *t*-ratios close to or in excess of conventional threshold levels. The switch from a negative to a positive effect for LFGRT as a cohort ages both is compatible with our model and occurs at a reasonable level of experience, approximately 10–20 years. In addition, the size of the effects is substantial. With the (negative) mean values of LFGRT for the five most recent cohorts, the estimated elasticities are approximately .20 (.10) for annual (weekly) earnings during years 1–5 of labor force experience and −.25 (−.15) during years 25–30 of labor force experience.

As with the estimates in table 3, the experience profiles in table 5 are U-shaped in all cases save one and thus provide additional support for our assumption of a weak correlation between initial earnings capacity and the propensity to invest in human capital.[23] The estimated variances of these coefficients are generally higher than the corresponding variances in table 3 as would be expected from the inclusion of LFGRT × experience.[24] The educational differences and the effects of unemployment are similar to those previously obtained.

---

[21] The use of alternative BLS labor force projections for the period 1981–2000 also failed to modify substantially the regression results reported in table 5. See the App. for a further discussion of labor force projections.

[22] For purposes of testing the predictions advanced in this paper, we believe that LFGRT is superior to the cohort size measure used by Welch (1979). Our variable can be measured for each cohort early in the work life when investment decisions are made. For most cohorts in our sample, this occurred outside the sample period. As a test of our LFGRT variable, we used it to estimate the model of mean cohort log earnings estimated by Welch (1979). The data set and specification were very similar. However, our CPS data extend through 1978 (rather than 1975), and we replaced Welch's cohort size measure with LFGRT. The results were qualitatively quite similar to those previously found. Negative values of LFGRT, which characterize the baby boom cohorts, have the effect both of lowering the intercepts and of steepening the slopes of log earnings profiles.

[23] All experience profiles are U-shaped when the "uncensored" estimates of the dependent variable are used (see n. 8).

An important question concerns the extent to which differences in LFGRT can account for the recent increase in within-cohort earnings inequality as reflected in the trend coefficients in table 3. In table 5, the trend coefficients for annual and weekly earnings obtained with the full sample are 25 percent and 40 percent lower, respectively, than the corresponding estimates in table 3. Similarly, the estimated trends for those with 11 or less and 13–15 years of schooling are at least 25 percent lower after the inclusion of LFGRT. For those with 12 years of education, however, the trend coefficients in table 5 are still positive, significant, and of essentially the same size as those in table 3. The education-specific regressions were also estimated with the measure of LFGRT based on the entire labor force. The principal change is that the resulting trend coefficients (not shown here) for annual and weekly earnings were at least 50 percent and 25 percent lower, respectively, than the trends in table 3 for all three of the lowest educational categories.

Table 5 shows that the effect of LFGRT for college graduates is consistent with our predictions. However, the absence of strong positive trends for this group constitutes an anomaly for which we have no ready explanation. As yet unidentified factors apparently served to maintain relatively constant within-cohort earnings inequality for these individuals over the sample period.[25]

In summary, the estimated effects of intercohort differ LFGRT are consistent with theoretical predictions and appea vide a partial explanation for recent increases in within-coho ings inequality. It is clear, though, that the growth in the dispe male earnings cannot be accounted for entirely by our empiric sures of the demographic influences emphasized in this study. future research, attention should be devoted to additional factors such as changes in transfer programs and taxes.

What of the future of earnings inequality? The implication of our findings is that recent and projected changes in labor force growth

---

[24] It should be noted that LFGRT is not strongly correlated with experience. The coefficient is 0.25 for the sample as a whole. Hence this factor cannot account for the "correct" results in table 5.

[25] Table 4 shows that college graduates had both a very recent (mid-1970s) peak in labor force growth rates and the smallest projected decline in growth rates from 1980 to 2000. On this basis, a smaller increase in the variance of log earnings for this group would be predicted. In addition, Freeman (1981), contrary to Welch (1979), reports that college graduates who entered the labor force in the early 1970s had slower rates of earnings growth than high school graduates, possibly indicating a smaller increase in investment. This is consistent with our finding of weaker trends in earnings inequality among college graduates; however, the evidence is clearly insufficient to resolve the issue at this stage.

TABLE 5

Regression Estimates of Determinants of Within-Cohort Variance of Annual and Weekly Log Male Earnings, 1967–78: The Impact of Changes in Labor Force Growth Rates

	Intercept and Schooling Dummies	Exper	Exper^2	Male Unemployment	Trend	Trend × Exper	Part-Time	LFGRT	LFGRT × Exper	R^2
Annual earnings.	.69 (31.7)	−.027 (9.0)	.00064 (9.1)	.014 (2.8)	.0070 (2.0)	−.00024 (1.7)		−120.0 (6.2)	9.7 (9.7)	.43
D12	−.27 (27.0)			...		...	...	...	...	...
D13–15	−.28 (28.0)					...	...	...	...	...
D16+	−.28 (28.0)	...	...	...	...	...	...	...	...	...
Weekly earnings.	.34 (16.7)	−.0019 (1.0)	.00008 (1.6)	−.006 (1.8)	.0035 (1.3)	−.00001 (.1)	.74 (12.5)	−33.8 (2.3)	1.8 (2.3)	.33
D12	−.13 (16.2)						...	...	...	...
D13–15	−.13 (13.0)	...				...	...	...	...	...
D16+	−.09 (9.0)	...	...	...			...	...	...	...

Education 11 years or less										
Annual earnings	79	− 037	00076	018	017	− 00035		−79.0	3.3	.44
	(11 7)	(10 3)	(12 7)	(2 3)	(2 8)	(1 9)		(1.8)	(2.1)	
Weekly earnings	45	− 011	00022	− 003	006	− 00001	62	−9 0	.23	.33
	(10.0)	(5 5)	(6)	(1 0)	(1)	(.1)	(8.3)	(.4)	(.3)	
Education: 12 years.										
Annual earnings	37	−.022	00057	011	015	− 00035		−25 2	4.2	39
	(5.9)	(5 5)	(7 1)	(2 2)	(5 0)	(2 7)		(1 0)	(3 2)	
Weekly earnings	.21	−.0036	00015	− 011	009	−.00010	61	−25.9	2.1	.28
	(4.6)	(1 2)	(2 5)	(3 1)	(4 5)	(1.1)	(7.7)	(1 5)	(2.3)	
Education. 13–15 years										
Annual earnings	37	− 013	.00046	019	004	− 00029		−74.9	4.4	21
	(6 0)	(1 9)	(2 9)	(1 7)	(6)	(.8)		(3 4)	(4.0)	
Weekly earnings	.18	0004	00012	− 0014	.0056	−.00015	52	−27.5	1 5	.17
	(3.9)	(.1)	(8)	(2)	(9)	(.6)	(3.5)	(1 6)	(1.8)	
Education: 16 years or more										
Annual earnings	39	− 021	.00063	014	− 009	.00034	. . .	−86.1	7.9	.13
	(7 7)	(3.0)	(3 2)	(1.1)	(1)	(7)		(2.4)	(3 2)	
Weekly earnings	18	− 0017	00025	− 0019	.0003	0001	1 31	−18.4	3.0	.14
	(3 3)	(3)	(8)	(.2)	(1)	(1)	(3 1)	(.7)	(1.4)	

NOTE.—Asymptotic normal statistics in parentheses.

rates temporarily increased the variance of log earnings among recent entrants. However, this effect should decline with experience and eventually become negative. Moreover, if completed fertility stays at or slightly below replacement, then the prospect is for a very gradual reduction (0.1 percent to 0.2 percent per decade) in the rate of growth of the adult population during the first half of the next century (U.S. Bureau of the Census 1977). Echoes of the post–World War II baby boom and bust will be felt, but the resulting swings in adult population growth rates will be minor in comparison with the original shocks. Hence, barring major shifts in immigration or labor force participation trends, the demographic sources of growth in earnings inequality that were the focus of this study should diminish in importance for future cohorts.

V. Conclusion

In this paper, we have analyzed the variance of log earnings within labor force cohorts of U.S. males, with particular attention to the influence of labor force growth rates. Estimates with data from the 1968–79 March current population surveys indicate that relative earnings inequality initially declines and then increases with labor force experience. The unemployment rate has a positive effect on the variance of annual log earnings but a negative, and much weaker, impact in the case of weekly earnings. Most important, evidence was provided of substantial increases in earnings inequality within labor force cohorts even after the levels of education, experience, and unemployment were controlled for. Such increases were most marked for, though not confined to, younger workers. Hence, recent growth in the dispersion of earnings among all males cannot be explained solely on the basis of widening earnings differentials among age groups or on the basis of recent growth in the proportion of young (high-variance) workers.

The model of human capital investment proposed by Ben-Porath was adapted to analyze differences among cohorts in the life-cycle profile of the variance of log earnings. It was shown that a rise in the growth rate for the human capital rental rate will stimulate investment and also increase the variance of postschooling investment and log earnings early in the life cycle. This positive effect on relative earnings inequality should decline with experience and possibly become negative. It was further shown that recent labor force entrants have faced just such a prospect of rapidly growing rental rates, because of the marked increase in labor force growth rates during the 1960s and 1970s and the equally dramatic deceleration in labor force growth expected during the remainder of this century. Therein lies a possible explanation for the recent growth in within-cohort earnings inequality.

The model was tested using estimates and projections of labor force growth rates for the period 1920–2000. Each cohort was assigned a measure of the expected changes in the rate of labor force growth over the first 20 years of experience. The estimated effects of this variable generally confirmed theoretical predictions and yielded sizable elasticities. Moreover, the foregoing strategy frequently produced substantial reductions in estimated time trends for the within-cohort variance of log earnings. Hence, it appears that recent increases in male earnings inequality may have been caused in part by the post–World War II baby boom and baby bust acting through the unprecedented large swings in labor force growth rates during the last half of the twentieth century. If replacement fertility rates are maintained, the resulting gradual deceleration in adult population growth during the first part of the twenty-first century will reduce significantly the importance of this demographic source of variation in earnings inequality.

Appendix

Estimates and Projections of the Labor Force

Our analysis required estimates and projections of the civilian
ages 16–64, from 1920 to 2000, by level of education. We firs
data and methods used to estimate labor force size irrespective
and then turn to the educational distribution.

Lebergott (1964, p. 512) provides estimates of the civilian lab
14 and over for the period 1920–46. From these figures, we su
mates of the number of participants ages 14–15 and 65 and ove
were derived by linearly interpolating between estimates from the
1930, and 1940 censuses of the population and the 1947 CPS. Annual esti-
mates for the period 1947–79 are from the *Employment and Training Report of the President* (U.S. Department of Labor 1980). The estimate for 1980 was taken from *Employment and Earnings* (U.S. Department of Labor 1981*b*)

Civilian labor force growth was clearly affected by World War II and to a lesser extent by the Korean War. Since these wars could not be anticipated by those who entered the labor force some 10–20 years prior to their occurrence, the effects of these conflicts were purged from the data by linearly interpolating between the years 1941–47 and 1950–54.

Howard Fullerton of the Bureau of Labor Statistics (BLS) kindly provided us with the most recent labor force projections for the period 1981–2000. Three sets of projections were available but did not produce substantially different estimates of the regression model for the variance of log earnings. Table 5 above contains the estimates obtained with the medium growth projections of the BLS.

Estimates and projections of the proportion of the civilian labor force with 11 or less, 12, 13–15, and 16 or more years of schooling were derived in the following manner: special Labor Force Report no. 240 (U.S. Department of Labor 1981*a*) provides estimates for the civilian labor force ages 18–64 in 1940, 1952, 1957, 1959, 1962, and 1964–79. Linear interpolation was used to derive estimates between the foregoing years.

Folger and Nam (1967) provide estimates for the educational distribution of the population ages 25 and over in 1920 and 1930. We derived estimates for the educational distribution of the labor force in 1920 and 1930 by adjusting these figures. The adjustment factor was the ratio *in 1940* of the proportion of the population 25 and over with a given level of education to the proportion of the civilian labor force ages 18–64 with the same level of education. Linear interpolation provided the estimates from 1920 to 1930 and from 1930 to 1940.

The most recent BLS projections of the educational distribution of the labor force are presented in Johnston (1973). These projections were for the years 1980, 1985, and 1990. First, these figures were adjusted for the observed forecast error over the period 1970–80. The adjustment factor was the ratio of actual to projected (in Johnston) average annual rates of growth. Next, a quadratic time trend was estimated for the number of labor force participants in each educational category using estimates and revised projections for the period 1970–90. Finally, these time trends were used to predict the educational distribution from 1990 to 2000, that is, the proportion of individuals in each category. The BLS projections provided by Fullerton were used to predict the total number of participants.

References

Ben-Porath, Yoram. "The Production of Human Capital and the Life Cycle of Earnings." *J.P.E.* 75, no. 4, pt. 1 (August 1967): 352–65.

Burmeister, Edwin. *Capital Theory and Dynamics.* Cambridge: Cambridge Univ. Press, 1980.

Chiswick, Barry R., and Mincer, Jacob A. "Time-Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985." *J.P.E.* 80, no. 3, pt. 2 (May/June 1972): S34–S66.

Dooley, Martin D. "Estimation in Censored Samples When There Is Heteroscedasticity." *Econ. Letters* (in press).

Folger, John K., and Nam, Charles B. *Education of the American Population.* Washington: Government Printing Office, 1967.

Freeman, Richard B. "The Effect of Demographic Factors on Age-Earnings Profiles." *J. Human Resources* 14 (Summer 1979): 289–318.

———. "Implications of the Changing U.S. Labor Market for Higher Education." Working Paper no. 697. Cambridge, Mass.: Nat. Bur. Econ. Res., June 1981.

Hause, John C. "The Fine Structure of Earnings and the On-the-Job Training Hypothesis." *Econometrica* 48 (May 1980): 1013–29.

Heckman, James J. "A Life-Cycle Model of Earnings, Learning, and Consumption." *J.P.E.* 84, no. 4, pt. 2 (August 1976): S11–S44.

———. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (January 1979): 153–61.

Henle, Peter, and Ryscavage, Paul. "The Distribution of Earned Income among Men and Women, 1958–77." *Monthly Labor Rev.* 103 (April 1980): 3–10.

Hurd, Michael. "Estimation in Truncated Samples When There Is Heteroscedasticity." *J. Econometrics* 11 (October/December 1979): 247–58.

Johnston, Denis F. "Education of Workers: Projections to 1990." *Monthly Labor Rev.* 96 (November 1973): 22–31.

Lebergott, Stanley. *Manpower in Economic Growth: The American Record since 1800.* New York: McGraw-Hill, 1964.

Mincer, Jacob A. *Schooling, Experience, and Earnings.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1974.

Paglin, Morton. "The Measurement and Trend of Inequality: A Basic Revision." *A E.R.* 65 (September 1975): 598–609.

Plotnick, Robert. "Trends in Male Earnings Inequality." *Southern Econ. J.* 48 (January 1982): 724–32.

Razin, Assaf. "Investment in Human Capital and Economic Growth." *Metroeconomica* 24 (May-August 1972): 101–16.

Schultz, T. Paul. "Long-Term Change in Personal Income Distribution. Theoretical Approaches, Evidence, and Explanations." In *The "Inequality" Controversy: Schooling and Distributive Justice,* edited by Donald M. Levine and Mary Jo Bane. New York: Basic Books, 1975.

Smith, James P., and Welch, Finis. "Inequality: Race Differences in the Distribution of Earnings." *Internat. Econ. Rev.* 20 (June 1979): 515–26.

Solow, Robert M. *Growth Theory: An Exposition.* New York: Oxford Univ Press, 1970

Stafford, Frank, and Duncan, Greg J. "The Use of Time and Technology by Households in the United States." In *Research in Labor Economics,* vol. 3, edited by Ronald G. Ehrenberg. Greenwich, Conn.: JAI, 1980.

U.S. Bureau of the Census. *Projection of the Population of the U.S., 1977–2050* Current Population Reports, ser. P-25, no. 704. Washington: Government Printing Office, May 1977.

U.S. Department of Labor. *Employment and Training Report of the President, 1980.* Washington: Government Printing Office, 1980.

———. *Educational Attainment of Workers, March 1979.* Special La Report no. 240. Washington: Government Printing Office, 198

———. *Employment and Earnings.* Washington: Government Print January 1981. (*b*)

Welch, Finis. "Effects of Cohort Size on Earnings: The Baby Boo Financial Bust." *J P.E* 87, no. 5, pt. 2 (October 1979): S65–S97

Welch, Finis, and Gould, William. "An Experience Imputation or a tion Experience." Unpublished paper. Los Angeles: UCLA, Octo

Zsigmond, S.; Picot, G.; Clark, W.; and Devereaux, M. S. *Out of School - the Labour Force.* Occasional Paper 81-570E. Ottawa: Ministry Indus., Trade, and Commerce, Statis. Canada, August 1978.

Gold Monetization and Gold Discipline

Robert P. Flood

Northwestern University

Peter M. Garber

University of Rochester

The substantial U.S. inflations of the 1970s led to some popular support for commodity-based money. Gold has had a special role in historical commodity-money schemes, and it emerged as a leading contender in recent discussions. In October 1980 Congress established the Gold Commission to study the possible remonetization of gold. In this paper we analyze some gold monetization schemes proposed to the commission. We find that the adoption of these proposals need not lead to the price-level stability their proponents seek. Even a well-designed commodity-money scheme is a foolproof inflation guard only when the scheme's permanence is guaranteed. Permanence may possibly be guaranteed by an underlying political economy that abhors inflation, but merely the enactment of a new ephemeral rule does not ensure permanence.

Stephen Salant's extensive comments proved enormously useful in developing this paper. Helpful comments also were given by Dale Henderson, Jeff Frankel, Ken Rogoff, and various participants in seminars at Rice University, the University of Florida, the University of Chicago, the University of Pennsylvania, M.I.T., the University of California, Berkeley, the Board of Governors of the Federal Reserve, the University of South Carolina, the Wingspread Conference, Harvard University, and Northwestern University. The paper was partly developed while Flood was a staff member of the Board of Governors of the Federal Reserve. The views presented here are our own and do not necessarily represent the views of the governors of the Federal Reserve or other members of their staff. We would like to thank the National Science Foundation for financial support.

[*Journal of Political Economy*, 1984, vol. 92, no. 1]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0006$01.50

Gold and its price have emerged as frequent topics of academic and government debate.[1] Public attention, focused on the gold market since the price of gold began fluctuating violently in 1980 and 1981, has turned to gold as a possible means of removing some discretion inherent in the current fiat money system. Since lack of adherence to monetary rules can lead to problems of dynamic policy inconsistency, academic economists have begun seriously to reconsider monetary standards based either on gold or on some other commodity as politically feasible methods of establishing rules.[2]

The political and intellectual resurrection of gold has coincided with the advent of three lines of thought concerning government gold market policy. First, Salant and Henderson (1978) have developed a partial equilibrium gold market model to analyze the interactions among rapid gold price rises, government gold auctions, and speculative attacks on government gold stocks. Salant (1983) places the possibility of a speculative attack on price stabilization schemes in a stochastic setting. Second, Barro (1979) has studied money and price dynamics under a version of the historical gold standard. Finally, others such as Laffer (1979) and Lehrman (1980, 1981) have proposed a new monetary gold standard under which the government would maintain a fixed nominal gold price while retaining the freedom to issue money without fixed gold backing.

In this paper we intend to provide a dynamic framework sim for analyzing various gold monetization policies. We m ideas in Salant and Henderson with those in Barro t model of a monetary system based on any storable comm

Besides analyzing a successful gold monetization, we st sibility of an unsuccessful gold standard (i.e., a goven plementation of a gold standard destined to collapse mediately or after a finite interval). Our results on the co gold standard draw on the work of Salant and Henderson (19 Krugman (1979), and Salant (1983). Finally, our analytical framework operates as a systematic means of categorizing any commodity standard's dynamic evolution. Using a simple graphical technique and given the policy environment, one can determine the nature of the potential equilibria during a gold standard.

We organize the paper into three sections. In Section I we introduce our model's basic components and study the paths of the relative gold price and the price level prior to a government's announcement that gold's nominal price will be fixed. In Section II we analyze the dynamics of the money and gold markets after the nominal gold price has been fixed. We find that equilibrium is not unique for some fixed gold prices. The market may determine equilibria consistent either with a permanently maintained gold standard or with a gold standard that eventually collapses. In Section III we examine the significance of the "discipline of the gold standard"; we demonstrate that the notion may lack operational meaning if the alternative fiat money system is not also "disciplined." Section III is followed by some concluding remarks.

---

[1] Congress decided to assess gold's monetary role by establishing a Gold Commission. While the commission was generally hostile to monetization of gold, its existence indicates that a return to a commodity base is a serious possibility

[2] For discussions of the time inconsistency problem, see Kydland and Prescott (1977) or Barro and Gordon (1983).

I. The World prior to Gold Monetization

To study the effects of gold price fixing, we first characterize the economy before the scheme is announced. We will use concepts and methods very similar to those of Salant and Henderson (1978). However, no government gold auctions will occur, and we will modify their model to suit our problem. Specifically, we introduce a monetary sector, modify the nature of gold demand for private uses, and analyze a policy of nominal gold price pegging. Throughout the paper we employ a continuous time model in which agents have perfect foresight.[3] We present our ideas in the form of a linearized example to make the concepts involved as concrete as possible. We divide our economy into two explicit sectors, a gold sector and a monetary sector. The real interest rate, ρ, is fixed at a constant level.[4]

The Gold Sector

Equations (1)–(3) partly describe the gold market's operation:

$$I = D(t) + G(t) + R(t), \tag{1}$$

$$\dot{D}(t) = v\left[\frac{\mu}{q(t)} - D(t)\right], \quad \mu, v > 0, \tag{2}$$

$$\begin{aligned} \dot{q}(t) &= \rho q(t), && \text{for } G(t) > 0, \\ &= 0, && \text{for } G(t) = 0. \end{aligned} \tag{3}$$

---

[3] However, the switch to the gold standard itself will be treated as a complete surprise. The perfect-foresight assumption applies to the dynamics of prices *after* the advent of the gold standard.

[4] Allowing the real interest rate to vary would complicate our analysis. We presume that the real side of our model consists of gold and a storable good that can be either consumed or stored to produce more goods. One unit of the stored consumable produces more consumables at the rate ρ, given exogenously as in Salant and Henderson (1978). Barro (1979) assumes that demand and supply functions are explicitly dependent on ρ, however, he exogenously fixes ρ to develop his gold standard's dynamics.

I, $D(t)$, $G(t)$, and $R(t)$ are the fixed total world gold stock, the quantity of gold at time t used in consumption and industry, the quantity of gold held privately as ingots or coins in speculative hoards, and the quantity of gold in government reserves, respectively. Until gold is monetized, we will assume $R(t)$ to be the positive constant $\bar{R}$.[5]

Equation (2) describes the law of motion for the total gold stock in consumption and industrial use; $\dot{D}(t)$ equals current consumption plus industrial gold purchases. When $\dot{D}(t) > 0$, speculators reduce their hoards to accommodate the flow. The quantity $\mu/q(t)$ is a desired stock of gold used in industry and consumption, where $q(t)$ is the relative price of gold in terms of other goods at time t. Central in generating any postmonetization price-level dynamics, the form of (2) reflects a slow transformation of gold into and out of consumption and industrial uses. Equation (2) is similar to a gold sector equation set up by Barro (1979).[6] However, we ignore gold mining and the effect that capital gains on gold may have on demand for consumption and industrial use.[7] Also, we treat gold as a perfectly durable good, which allows the possibility that gold may be disgorged from other uses into hoards when the relative gold price rises sufficiently.[8]

Equation (3) requires that gold's relative price must rise at the real interest rate while private speculators hold gold. If the relative pri were expected to rise faster than ρ, speculators would bid current price until $\dot{q}(t)/q(t)$ fell to ρ. If the price were expe more slowly than ρ, then speculators would drive the cu downward.[9]

Equations (1)–(3) form a differential equation system i $q(t)$. Equation (3) yields

$$q(t) = q(0) \exp (\rho t)$$

---

[5] These assumptions preclude any gold-mining activity, a specification similar to that of Salant and Henderson (1978)

[6] Surprisingly, Barro (1979) worked out the dynamics of his gold standard model using a static expectations assumption. This produced a stable dynamic system that is comparable only in its steady state to the system developed below

[7] In an early version of this paper, we solved the model for a case in which capital gains enter into desired gold use Because no substantive result depends on this assumption, we exclude it here to simplify our report

[8] While our methods are similar to those of Salant and Henderson (1978), our gold market model differs from theirs in some important respects Salant and Henderson interpret the variable that, in the context of their paper, is analogous to our $D(t)$ as a demand for a perishable good Since this gold demand must always be nonnegative in their paper, we have avoided using their form here. Otherwise, we would preclude the possibility of a postmonetization gold inflation and the disgorging of consumer gold that occurred in 1980 and 1981 Also, we assume that the flow gold demand depends negatively on $D(t)$, the existing gold stock already used in consumption and industrial forms. This may reflect diminishing marginal productivity associated with additional industrial gold uses.

[9] Salant and Henderson (1978) discuss the speculative pricing of such resources in more detail.

when $G(t) > 0$. Substituting from (4) into (2) generates a differential equation in $D(t)$ whose solution is[10]

$$D(t) = D(0) \exp(-vt) + v \exp(-vt) \int_0^t \left[\frac{\mu}{q(0)}\right] \exp[(v - \rho)u] du, \tag{5}$$

where $D(0)$ is the initial value of $D(t)$.

To complete the solution we must determine $q(0)$. Since speculators hoard gold only in anticipation of future consumption and industrial uses, the speculative gold stock must be exhausted when additional gold use in these areas ceases. If T is the date when $\dot{D}(t) = 0$, then $G(T)$ must equal zero. Since $G(T) = 0$, equation (1) implies

$$I - \bar{R} = D(T). \tag{6}$$

Combining equations (6) and (2) and rearranging, we find

$$q(T) = \frac{\mu}{I - \bar{R}}. \tag{7}$$

Equation (7) indicates the "choke price" for new consumption and industrial gold use, the terminal condition required to solve for $q(0)$. To find $q(0)$, we employ (7) and (4) to yield

$$q(0) = \left(\frac{\mu}{I - \bar{R}}\right) \exp(-\rho T). \tag{8}$$

Our solution for the gold sector will be complete once we have determined the unknown choke date, T, which is a function of the initial condition $D(0)$. T is determined by substituting from (8) into (5) for $q(0)$ and setting $D(T) = I - \bar{R}$. This sequence of steps produces a single equation in T whose solution can be characterized as $T = f[D(0)/(I - \bar{R})]$ with $f' < 0$, $f(1) = 0$, and $f(0) = \log(\rho/v)/(\rho - v) \equiv k > 0$.[11]

Determining $q(0)$ completes our solution for the gold sector. Given $q(0)$, equation (1) indicates the time path of $q(t)$ from 0 to T, and $q(t) = q(T)$ for $t > T$. Given the initial condition $D(0)$, equation (5) determines the time path of $D(t)$ from 0 to T with $D(t) = D(T) = I - \bar{R}$ for $t > T$. The gold sector alone determines gold's relative price path and the quantity of gold held in speculative hoards. To find the nominal gold price, however, we must append a monetary sector.

---

[10] The differential equation is $\dot{D}(t) = -vD(t) + [v\mu/q(0)] \exp(-\rho t)$, when $G(t) > 0$.

[11] The function determining T at time zero is $1 = D(0) \exp(-vT)/(I - \bar{R}) + v\{1 - \exp[-(\rho - v)T]\}/(v - \rho)$. For the special case of $D(0) = 0$, the solution for T is $T = \log(\rho/v)/(\rho - v)$, for $\rho \neq v$.

The Monetary Sector

The operation of the monetary sector has no real effects in the model prior to gold monetization. We introduce the monetary sector now as a preliminary to the analysis of Section II.

The following equations govern money market equilibrium:

$$\frac{M(t)}{P(t)} = \beta - \alpha i(t), \quad \beta, \alpha > 0; \tag{9}$$

$$i(t) = \rho + \frac{\dot{P}(t)}{P(t)}. \tag{10}$$

$M(t)$, $P(t)$, $i(t)$, and $\dot{P}(t)/P(t)$ are the nominal quantity of money, the aggregate price level, the nominal rate of interest, and the expected and actual rate of inflation, respectively. Equation (9) is a linearization that is appropriate for $\beta - \alpha i(t) > 0$ and for $\rho + [\dot{P}(t)/P(t)] > 0$.

In this section we assume that $M(t)$ equals the quantity of fiat currency in circulation. We also assume that the government prints fiat money at a constant positive rate π to finance a fixed positive level of real expenditure g. Since $\dot{M}(t)/P(t) = g$, it follows that $\pi = gP(t)/M(t)$. Money market equilibrium then requires that $\dot{P}(t)/P(t) = \pi$, a condition used in (9) and (10) to obtain

$$P(t) = \frac{M(t)}{\phi - \alpha\pi},$$

where $\phi = \beta - \alpha\rho > 0$. We define $Q(t)$ as the nominal gol
$q(t) = Q(t)/P(t)$.[12]

This model lacks interesting interactions between the g
monetary sectors. However, such interactions are introd
policy of gold monetization, the topic of the next section

II. Equilibrium Dynamics under a Gold Standard

We now study the price level, money market, and gold market dynamics after a government fixes gold's nominal price, committing itself to buy and sell gold at some price $\bar{Q}$. We assume the price fixing occurs immediately and completely surprises the private sector. Initially, we assume that the government does not sterilize gold market interventions in other markets. Consequently, government gold market operations affect the money stock. Through this channel a government's decision to fix gold's price links the money and gold markets. A price-fixing scheme may be permanently viable, temporarily viable, or nonviable. A temporarily viable scheme will function for some finite period, collapsing at a predictable time. A nonviable gold standard will collapse at the instant of its implementation. We will develop a general framework capable of determining the dynamics of a particular gold standard.

---

[12] Since $P(t)$ contains the rental price of gold in consumption form, our relative price $q(t) = Q(t)/P(t)$ is not equal to the price of gold divided by the price of goods other than gold. However, we may assume that gold's consumption rental price is proportional to gold's price and that $P(t)$ is a geometric average of gold's rental price and the price of other goods. Then our $q(t)$ is proportional to the price of gold divided by the price of goods other than gold raised to the power $(1 + \Omega)$, where Ω is gold's share in the price index. In our analysis we assume that Ω is sufficiently small that it may be ignored without consequence. However, if Ω were large then our eq. (3) would become $q(t) = \rho(1 - \Omega)q(t)$. This change would not alter our results substantively

The Shadow Free Market Gold Price

We define the shadow free market gold price $\tilde{Q}(t)$ as the equilibrium nominal gold price that would prevail if speculators purchased all of the government's gold reserves at time t, paying $\bar{Q}$ per unit. Such an attack would force gold demonetization, allowing the market freely to determine the gold price $\tilde{Q}(t)$. A gold standard remains viable while the shadow free market gold price $\tilde{Q}(t)$ does not exceed the government's fixed price $\bar{Q}$. If $\tilde{Q}(t) > \bar{Q}$, speculators will profit by attacking the gold standard. If $\tilde{Q}(t) < \bar{Q}$, an attack on the government's gold stock would generate speculative losses. When $\tilde{Q}(t) = \bar{Q}$ an attack produces neither profit nor loss. Since our analysis depends crucially on the shadow price $\tilde{Q}(t)$, we will develop the concept in greater detail.

Determining $\tilde{Q}(t)$ requires some assumptions about government monetary policy during and after a gold standard epoch. We assume that during a gold standard the government undertakes open market operations only in the gold market. The government will back any additional monetary issue 100 percent by gold purchases and match any gold reserve contraction by an equally valued monetary contraction. If the gold standard collapses, we assume that the government immediately resumes printing money at the positive rate π. Of course, many different assumptions about monetary policy may be analyzed in our model, and below we discuss some alternatives.[13]

To determine $\tilde{Q}(t)$ we find $\tilde{P}(t)$ and $\tilde{q}(t)$, the shadow postcollapse price level and relative gold price, respectively. We define time 0 as the first moment that the gold standard functions. Under the government's monetary policy, the money supply will be $M(t) = M(0) - \bar{Q}\bar{R}$

---

[13] Our assumption about the gold standard monetary policy requires that the government either develop a nonmonetary method of financing g or reduce spending by g. In Sec. II we discuss the implications of an alternative assumption about postcollapse money growth.

if a collapse occurs at any time $t \geqslant 0$. The postcollapse price level will be

$$\tilde{P}(t) = \frac{M(0) - \bar{Q}\bar{R}}{\phi - \alpha\pi}. \tag{12}$$

For now, we assume that $[M(0) - \bar{Q}\bar{R}] > 0$, implying that the existing money stock can purchase all the government's initial gold reserves.

To determine the postcollapse relative gold price $\tilde{q}(t)$, we use the previous section's gold market model. Following an attack that exhausts government reserves, private agents hold the entire gold supply only for productive use or in speculative hoards. When hoards are exhausted at time T, the relative gold price will attain $q(T) = \mu/I$.[14] The immediate postcollapse relative price is $\tilde{q}(t) = (\mu/I)\exp[-\rho(T - t)]$. The amount of time $(T - t)$ until the choke date depends on the initial level of $D(t)$ and the terminal level I. In particular, $(T - t) = f[D(t)/I]$, with $f' < 0$, $f(1) = 0$, and $f(0) = k > 0$ as in Section I. Combining these results, we can express $\tilde{Q}(t)$ as

$$\tilde{Q}(t) = \tilde{P}(t)\tilde{q}(t) = \frac{[M(0) - \bar{Q}\bar{R}]\mu}{(\phi - \alpha\pi)I}\exp\left\{-\rho f\left[\frac{D(t)}{I}\right]\right\}. \tag{13}$$

This expression provides an important building block used extensively below.

Gold Standard Dynamics

If the government sets $\bar{Q} > \tilde{Q}(0)$, the gold standard will least temporarily because no profit will arise from an imm ulative attack. We will now analyze the dynamics of a wor dard.

The money supply provides the key element integrating the m tary and gold sectors after price fixing. After gold monetization the money supply is

$$M(t) = M(0) + \bar{Q}[R(t) - \bar{R}] + \bar{Q}G(t). \tag{14}$$

$M(0)$ is the quantity of money existing just prior to gold monetization; $\bar{Q}[R(t) - \bar{R}]$ is the nominal value of government gold purchases after the price is fixed; $\bar{Q}G(t)$ is the nominal value of gold used directly as money. Since gold's price is fixed in terms of currency, gold will yield the same nominal capital gain as currency. Gold can also serve as money, so those hoards remaining in private hands must, as perfect substitutes for currency, constitute part of the money stock. Since we cannot determine the division of monetary gold between government and private holding, it is convenient to define

---

[14] The postcollapse choke date T is not the same as the choke date discussed in Sec. I since the total gold stock available for consumption increases by the government's reserves. However, the logic used in Sec. I to determine T is also applicable in this case.

$$X(t) = R(t) + G(t). \tag{15}$$

Our framework allows us to keep track of $X(t)$ but not of its components.

Using the variable $X(t)$, we can write the money supply as

$$M(t) = M(0) - \bar{Q}\bar{R} + \bar{Q}X(t). \tag{16}$$

Since we assume that the government's only open market operations occur in the gold market, only movements in $X(t)$ affect $M(t)$ after gold's price is fixed. Substituting from equation (16) for the money supply in (9), we obtain one of the equations of motion for the gold standard regime,

$$\dot{P}(t) = \left[\left(\frac{\phi}{\alpha}\right)P(t)\right] - \left[\left(\frac{\bar{Q}}{\alpha}\right)X(t)\right] - \frac{M(0) - \bar{Q}\bar{R}}{\alpha}. \tag{17}$$

The dynamics of the variable $X(t)$ are determined from those of $D(t)$. Since $D(t) = [I - X(t)]$ and $\dot{X}(t) = -\dot{D}(t)$, we substitute for $D(t)$ and $\dot{D}(t)$ in (2) to derive

$$\dot{X}(t) = -\left(\frac{v\mu}{\bar{Q}}\right)P(t) - vX(t) + vI. \tag{18}$$

Equations (17) and (18) are the laws of motion of $P(t)$ and $X(t)$ under the gold standard.

Setting $\dot{X}(t) = \dot{P}(t) = 0$ in (17) and (18), we find the steady state ($\hat{P}$, $\hat{X}$) of the gold standard:

$$\hat{P} = \frac{\bar{Q}(I - \bar{R}) + M(0)}{\phi + \mu} \tag{19}$$

$$\hat{X} = \frac{\phi I + \mu\{\bar{R} - [M(0)/\bar{Q}]\}}{\phi + \mu}. \tag{20}$$

$\hat{P}$ and $\hat{X}$ serve as terminal values for $P(t)$ and $X(t)$ if $\tilde{Q}(t) < \bar{Q}$ for all t. However, if $\tilde{Q}(t) > \bar{Q}$ for some t, another terminal condition will prevail. We next examine jointly the dynamics of the gold standard and of $\tilde{Q}(t)$ to determine equilibrium paths for different values of $\bar{Q}$.

Terminal Conditions and Equilibrium Paths

Figure 1 provides a systematic means of categorizing equilibrium paths and terminal conditions under a gold standard. We plot values of $\bar{Q}$ and $X(t)/I$ on the vertical and horizontal axes, respectively. The

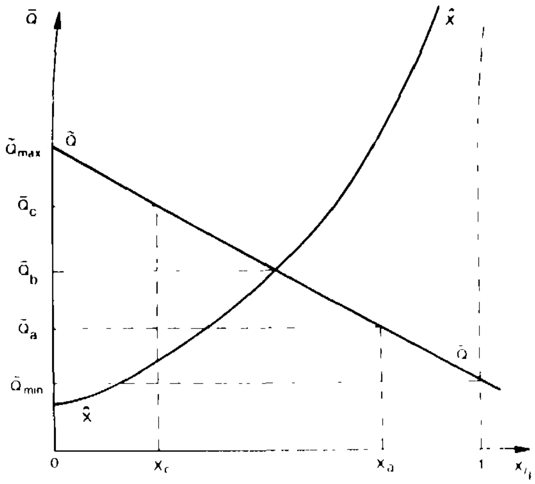


FIG. 1

line $\tilde{Q}\tilde{Q}$ is a locus of points $(X, \bar{Q})$ derived from equation (13) by setting $\tilde{Q}(t) = \bar{Q}$ and using $D(t)/I = 1 - [X(t)/I]$. For points above the line, $\bar{Q} > \tilde{Q}(t)$, while for points below the line $\bar{Q} < \tilde{Q}(t)$. $\tilde{Q}\tilde{Q}$ slopes negatively with vertical intercept $\tilde{Q}_{max} = M(0)\mu/[(\phi - \alpha\pi)I + \bar{R}\mu]$. When $X/I = 1$, the height of $\tilde{Q}\tilde{Q}$ is $\tilde{Q}_{min} = \tilde{Q}_{max} \exp(-\rho k)$.

We also plot the line $\hat{X}\hat{X}$ from equation (20), which produces steady-state $\hat{X}$ values as a function of $\bar{Q}$. $\hat{X}\hat{X}$ slopes positively with a verti intercept at $\mu M(0)/(\phi I + \mu\bar{R})$. As $\bar{Q}$ approaches infinity proaches $[(\phi + \mu\bar{R})/I]/(\phi + \mu) \leq 1$.

The $\tilde{Q}\tilde{Q}$ locus divides figure 1 into three regions: (i) $\bar{Q} <$ $\tilde{Q}_{max} \geq \bar{Q} \geq \tilde{Q}_{min}$, and (iii) $\bar{Q} > \tilde{Q}_{max}$. We will initially examin and third.

A Nonviable Gold Standard

If a government tries to establish a $\bar{Q}$ less than $\tilde{Q}_{min}$, then $\bar{Q} < \tilde{Q}(t)$ for all X. An attack will occur immediately, and the gold standard will be stillborn.

A Permanently Viable Gold Standard

If the government establishes a $\bar{Q}$ greater than $\tilde{Q}_{max}$, then $\bar{Q} > \tilde{Q}(t)$ for all X and speculators will never attack the gold standard. In this region the appropriate terminal value of $X(t)$ is $\hat{X}$. The particular path followed to $\hat{X}$ depends on the initial $X(0)$ value.

In figure 2 we depict the dynamics implicit in (17) and (18) for a

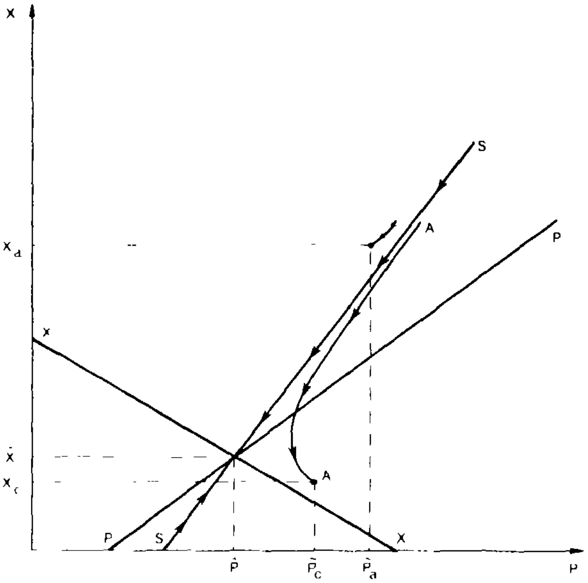


FIG. 2

permanent gold standard. The lines PP and XX are the loci of points $[P(t), X(t)]$ such that $\dot{P}(t) = 0$ and $\dot{X}(t) = 0$, respectively. PP and XX intersect at the dynamic system's steady-state price level and monetary gold stock, $\hat{P}$ and $\hat{X}$. We construct figure 2 for a particular value of $\bar{Q}$; an increase in $\bar{Q}$ would push the associated $(\hat{P}, \hat{X})$ to the northeast. The dynamics imply that the steady state is a saddle point with a unique stable branch, indicated by the line SS.[15] For a permanently viable gold standard we assume that economic forces determine $P(t)$, given $X(t)$, such that $P(t)$ and $X(t)$ lie on the line SS.[16] For example, if $X(0) > \hat{X}$, then $P(0) > \hat{P}$ and both $P(t)$ and $X(t)$ will move along SS approaching the steady state.

Transient Gold Standards and Multiple Equilibria

We now examine the dynamics associated with different values of $\bar{Q}$ such that $\bar{Q}_{\min} \leq \bar{Q} \leq \bar{Q}_{\max}$. We define the pair $(X_b, \bar{Q}_b)$ as the intersection point of $\hat{X}\hat{X}$ and $\tilde{Q}\tilde{Q}$ in figure 1 and consider $\bar{Q}$ values less than, equal to, and greater than $\bar{Q}_b$.

---

[15] The equation of SS is $P(t) = AX(t) + \hat{P} - A\hat{X}$, where $A = \bar{Q}/(\phi - \alpha\lambda_1) > 0$ and λ_1 is the negative root of the equation system (17) and (18). To be precise, $\lambda_1 = \frac{1}{2}(-[\phi/(\alpha - v)] - \{[\phi/(\alpha - v)]^2 + 4(v/\alpha)(\phi + \mu)\}^{\frac{1}{2}}) < 0$. For later use we also record the positive root of the system, $\lambda_2 = [(-v/\alpha)(\phi + \mu)]/\lambda_1 > 0$.

[16] This assumption requires the absence of speculative bubbles, by which we mean that prices are determined entirely by market fundamentals. We impose this requirement throughout the paper.

A transient gold standard.—If $\tilde{Q}_{\min} \leq \bar{Q} < \bar{Q}_b$, then the gold standard's steady state is not attainable since $\tilde{Q}(t)$ exceeds $\bar{Q}$ at the steady state. In figure 1, $\bar{Q}_a$ represents such a $\bar{Q}$ setting. If the pair $[X(0), \bar{Q}]$ lies on or below $\tilde{Q}\tilde{Q}$, speculators will attack the gold standard immediately. If $[X(0), \bar{Q}]$ lies above $\tilde{Q}\tilde{Q}$, the gold standard will survive temporarily, but it must eventually collapse at a predictable time z.

The notion of a transient gold standard, implying a predictable collapse, may seem alien to the economist who senses foreseeable profit opportunities arising from discrete gold price and price-level movements. However, Salant and Henderson (1978), Krugman (1979), and Salant (1983) have shown that an anticipated collapse of a price-fixing scheme is barren of profit opportunities. In a gold standard collapse, the price level and gold accumulation rate will adjust immediately to preclude predictable speculative profit opportunities. Though it is an important development, the Salant and Henderson (1978) result has received scant attention; hence, we will spend some effort to present the concept in the context of our model.

It is possible to derive a method for determining z, the time of the collapse. However, since we were unable to construct a closed-form solution for z, we will present only a general analysis of the pr
continuity results necessary to approach a solution, leaving
plete discussion of the determination of z to the Appendix
$\hat{P}(z)$ are the postcollapse gold price and price level, respect
is the fixed gold price, the collapse will occur only if $\tilde{Q}($
agents foresee $\tilde{Q}(z) > \bar{Q}_a$, then they anticipate a profit opp
z. Competition among agents planning to gain when the
dard collapses will lead some agents to purchase the gov
gold stock just prior to z to reap the entire speculative profit. Since t
collapse would then occur before z, a foreseeable collapse at z can occur only if $\tilde{Q}(z) = \bar{Q}_a$.

Since goods are storable, no anticipated price-level jumps can coincide with the gold standard's collapse, although the money stock will discretely decline. Foreseeable price-level discontinuities would open profitable speculative opportunities involving the storable good. The equilibrium price level will be determined currently such that at time z its path is continuous.

We can graphically analyze the dynamics of a temporarily viable gold standard using figure 2. $\hat{P}_a$ represents the postcollapse price level associated with $\bar{Q}_a$;[17] X_a is the quantity of monetary gold when $\tilde{Q}(z) =$

---

[17] It can be shown that $\hat{P}_a$ exceeds $\hat{P}$ when $\bar{Q} \leq \tilde{Q}_{\min}$. We realize that $\hat{P}$ and $\hat{X}$ are both positive functions of $\bar{Q}$. However, we have used fig. 2 to analyze dynamics for several $\bar{Q}$ values to preserve simplicity of presentation. The reader should be reminded that the steady state will actually move as we examine different $\bar{Q}$ values.

$\bar{Q}_a$. Recognizing the coming collapse, agents foresee that afterward the price level and relative gold price will be driven by the dynamic laws established in Section I. Since $(\tilde{P}_a, X_a)$ rather than $(\hat{P}, \hat{X})$ serves as the terminal condition for the gold standard dynamics, the price level and monetary gold stock will not generally move along the saddle path during the gold standard's temporary existence.[18] Instead, the system will follow an ostensibly unstable path and collapse when the monetary gold stock reaches X_a, the gold hoard that agents are just willing to hold in the pure fiat system of Section I. In the Appendix we show how to compute $\tilde{P}_a$, $\tilde{Q}(z)$, X_a, and z.

Multiple equilibrium paths.—We now study the possible equilibria that arise when the fixed gold price equals $\bar{Q}_c \geqslant \bar{Q}_b$ in figure 1. If $X(0) < X_c$, the gold standard again will be immediately attacked. However, if the initial monetary gold stock exceeds or equals X_c, multiple equilibrium paths will arise.

The multiplicity depends purely on agents' expectations about the viability of the gold standard. If agents believe that the gold standard will function permanently, equilibrium will occur along the stable path SS associated with $\bar{Q}_c$ in figure 2, and $P(t)$ and $X(t)$ will decline toward the steady state $(\hat{P}, \hat{X})$. However, if agents believe that the gold standard will survive only temporarily, their behavior will establish the equilibrium along the path AA in figure 2. The terminal point for AA is $(\tilde{P}_c, X_c)$, where $\tilde{P}_c$ is the shadow price level associated with $\bar{Q}_c$. When the system attains this terminal point, $\tilde{Q} = \bar{Q}_c$ and the gold standard will be attacked.[19]

In our model the potential for multiple equilibria arises whenever the government extracts resources through money creation in a post-collapse fiat regime. Such a policy ensures that the $\dot{Q}\dot{Q}$ and $\dot{X}\dot{X}$ loci of figure 1 intersect in the positive quadrant. As g and π decline, the $\dot{Q}\dot{Q}$ locus shifts downward, intersecting the $\dot{X}\dot{X}$ locus on the vertical axis when $\pi = 0$. When $\pi = 0$, multiple equilibria will occur only for a fixed gold price given by the height of this intersection point.

Price-Level Effects of Gold Monetization

Since gold monetization is intended to provide a more stable price level than that of the current fiat system, we study the response of the price level to a newly implemented gold standard.

When the gold standard is immediately nonviable, the attack on the government's gold stock causes the money supply to fall discretely by an amount $\overline{Q}\overline{R}$. If the government does not otherwise alter its previous money growth plans, the price-level path is displaced downward by $\overline{Q}\overline{R}/[\phi - \alpha\pi]$.

---

[18] Since $\tilde{P}_a > \hat{P}$ and $X_a > \hat{X}$, the gold standard will terminate northeast of $(\hat{P}, \hat{X})$ in fig. 2. However, without precise parameter values, we do not know where $(\tilde{P}_a, X_a)$ lies with respect to SS.

[19] Since $\tilde{P}_c > \hat{P}$ and $X_c < \hat{X}$, the path approaching $(\tilde{P}_c, X_c)$ must lie to the right of SS.

If the gold standard is permanently or temporarily viable, its price-level effects depend on the specific monetary growth path followed before the gold standard is announced, the price of gold, and the model's initial conditions and parameters.

For example, we may consider the price-level effects of a permanently viable system. Prior to the implementation of the gold standard at time 0, agents had expected a growing money supply. When gold's price is fixed at $\overline{Q}$, the monetary aggregate immediately jumps from $M(0)$ to $M(0) + \overline{Q}G(0-)$, where $G(0-) > 0$ is the private gold hoard just before monetization. Once the gold standard is implemented, movements of $X(t)$ drive the money stock. Depending on the choice of $\overline{Q}$, the money stock may rise, decline, or remain static through time.

Since the price level is a weighted average of future money stocks, the gold standard need not stabilize prices relative to the fiat system.[20] In order to determine the price-level effects of the monetary system change, we must know the future money stock paths. Evidently, the money stock can evolve many different ways, generating gold inflation, gold deflation, or price stability. The gold standard may be more or less stabilizing than the fiat system, depending on the initial speculative gold stocks, the parameters of gold and money dem
the fixed gold price. In practice, the price level can be sta
sterilizing the shift in the money growth path at the instant o
zation and choosing a gold price placing the system at its ste

Some Extensions of the Analysis

At this point we wish to consider relaxing three important assum tions: (i) $\overline{Q}\overline{R} < M(0)$; (ii) the only open market operations undertaken while the gold standard functions are gold market operations; and (iii) the money growth path after a collapse is the pre–gold standard expected growth path.

If open market operations occur only in the gold market during the gold standard era, the assumption that $\overline{Q}\overline{R} < M(0)$ implies that $\overline{Q}R(t) < M(t)$ for all t. Enough currency to purchase all the government's gold at the fixed price $\overline{Q}$ always will exist, a necessary condition for a collapse. Alternatively, if we assume $\overline{Q}\overline{R} > M(0)$, the gold standard cannot collapse. When open market operations are allowed only in the gold market, further currency creation corresponds to additions to government gold stocks. Because the government's reserves, always maintaining a greater value than outstanding claims, will never be attacked, the analysis for a permanently functioning gold standard is applicable.

---

[20] For an arbitrary money growth path, money market equilibrium will generate $P(t) = \exp(\phi t/\alpha) \int_t^\infty [M(u)/\alpha] \exp(-\phi u/\alpha)du$

We may also relax the assumption that open market operations occur only in the gold market during the gold standard era. If gold's price exceeds the minimum, permanently viable price, the government, though continually maintaining the fixed gold price, possesses some discretion to increase the money stock through channels other than gold purchases. Unless it is expected that they will be transitory, such monetary expansions raise the shadow free market price, $\tilde{Q}(t)$. However, if the government limits the monetary expansion so that $\tilde{Q}$ always exceeds $\tilde{Q}_{max}$, it can still ensure a permanently functioning gold standard.

Finally, we have assumed that the money growth path following the collapse is identical to the pre–gold standard path, $\dot{M}(t)/M(t) = \pi$. This assumption is applicable if the government extracts a constant resource flow through money creation in a postcollapse economy. However, the government's postattack, shadow revenue requirements may continually rise during the gold standard's operation, so that the postattack shadow inflation rate also rises. Under such circumstances, the $\tilde{Q}\tilde{Q}$ locus in figure 1 would continually shift upward through time, forcing the eventual collapse of an otherwise well-functioning gold standard.

III. The Discipline of the Gold Standard

Any pressure to reestablish a commodity base for the monetary system rests on a conviction that "discipline" generated by such constraining rules yields beneficial price-level stability or predictability. However, viewed in a dynamic context of periodic movements from commodity to fiat monetary systems, the discipline of commodity systems is a mask obscuring the forces that drive the global dynamic money creation process. Since all commodity standard interludes eventually are superseded by fiat systems, the notion of discipline can characterize only the underlying forces generating the entire dynamic monetary process. The application of this concept to only one facet of a general policy sequence, overlooking the empirical realization that all previous commodity standards have collapsed into fiat systems, may lead to misplaced confidence in prescriptions to return to a commodity base.

Behind the sequential transitions from one monetary scheme to another that we have analyzed must lie a political economy that we have ignored. Such political economic forces determine the complete dynamic panorama of the monetary process. Their nature can be gauged only by the entire sequence of monetary bases, not by the ephemeral monetary base that currently may operate. If the fiat segments of the money creation process are inflationary enough, then a currently existing commodity standard may not be viable, even though if it continued the money stock would grow according to the commodity standard's dynamic laws. A commodity system can be interpreted as a discipline-imposing rule only if the commodity standard's permanence is somehow guaranteed. As there is no means to secure such permanence, the notion of a commodity standard as a stabilizing rule is a chimera. The monetary system is stable only if the global money generation scheme and agents' confidence in it produce stability, not if money is temporarily created according to some transitional rule.

If conditions arise that indicate the imminence of highly inflationary monetary policies contingent on a gold standard's collapse, the gold standard, though maintained by a government conditionally and scrupulously obeying its rules, may not survive. The gold standard is but one phase of an overall policy; lack of stability of the entire constellation of conditional monetary standards precludes the stability of any one of its commodity standard phases. In contrast, current policy discussions prescribe the adoption of some commodity standard *because* conditional fiat systems are too unstable. However, simply comparing the price-level dynamics of two permanent money g[illegible] schemes yields misleading results. The central political [illegible] problem, that fiat systems are unstable, can override any [illegible] properties of permanently viable commodity standards [illegible] political system somehow ensures stability of the cond[illegible] epochs of its money generation process, implementing a c[illegible] standard may be a provisional, if not momentary, exercise [illegible] duct an analysis of changes in monetary systems requires the recognition that all monetary generation rules are temporary, predictably to be replaced under some circumstances.

IV. Conclusion

Many possible commodities can serve as monetary standards. However, the sequence of items on which the monetary system has been based in modern times has been so limited that it can readily be characterized as a "gold-paper-gold" cycle. Since the current paper phase of the cycle has allowed price-level instability, there has been some agitation to return to the gold standard. Proponents of such a return presume that simply allowing the market to determine the price at which gold will be fixed is sufficient to produce an "appropri-

Matching, Turnover, and Unemployment

Boyan Jovanovic

New York University

This paper develops the view that workers move into and out of the labor market because of changes in the perceived value of their market opportunities. A search model is combined with a matching model to yield movement of workers from job to job and into and out of employment.

I. Introduction

Work histories tell us that workers move from job to job and that they also move into and out of employment. Yet theory tackles these two types of moves one at a time. Turnover models are silent on moves into and out of employment, while unemployment models leave moves from job to job unexplained. Since both kinds of moves can lead to large changes in welfare, surprisingly few attempts are made at building models that explain both.

One exception is Burdett (1978). In his model, search on and off the job leads to moves from job to job and from unemployment to employment. But it does not give rise to any moves from employment to unemployment: only the new arrivals in the labor market are unemployed.

The fact is, most of the unemployed have had jobs before. To get moves out of employment, Toikka (1976) and Flinn and Heckman (1982) assume fluctuations in people's value of nonmarket time. Changes in nonmarket opportunities thus give rise to moves out of the labor market.

I thank Kenneth Burdett for suggestions that inspired the writing of this paper. James Heckman, Glenn MacDonald, Charles Wilson, and two referees made useful comments on an earlier draft.

[*Journal of Political Economy*, 1984, vol. 92, no. 1]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0008$01.50

This paper develops an alternative view. This view is that workers move into and out of the labor market because of changes in the perceived value of their *market opportunities*. The paper takes a search model—much like Burdett's—and combines it with my matching model (Jovanovic 1979*a*). The result is a model in which all moves are endogenous; workers move from job to job, and they move into and out of employment.

II. The Model

The production technology is the same as in Jovanovic (1979*a*). Production functions have constant returns to scale, and labor is the only input. At each instant, the worker's output is observed by him and his employer; both are always equally informed about the quality of their match.

The quality of the match between the worker and his employer is denoted by μ; it stands for the *average* output produced over a period of time of unit length; deviations of actual output from its average level are normally distributed. If $X(t)$ is the worker's *cumulative* output over a period of length t, then $X(t) = \mu t + \sigma_x Z(t)$, where $\sigma_x > 0$ is a *known* constant, identical over all possible matches. Time is continuous, and $Z(t)$ is the standard Wiener process, $Z(t) \sim N(0, t)$ This process has independent increments—cov $Z(t), Z(t') =$ min Unknown at the start, the quality of the match, μ, reveals its match produces output over time.

Among all possible matches between workers and firms, σ_μ^2). With no other information, a worker and his employer quality of their match as having been drawn from this dis They update their beliefs when they get some evidence.

Evidence comes in two ways. The firm first screens the worker outcome of the screen is a "noisy" estimate of μ, $m = \mu + \epsilon$, where $\epsilon \sim N(0, \sigma_m^2)$. The smaller σ_m, the better the screen. This parameter limits how much one can learn about the match before one observes any of its output. If $\sigma_m = \infty$, a match is a pure *experience good*—nothing can be learned about the match before it produces output. But if $\sigma_m = 0$, the match is a pure *inspection good*—one can learn its quality merely by inspecting it.

The second source of evidence is direct observation of the output of the match. For a match with screen outcome m, job tenure t, and cumulative output $X(t) = x$, the posterior distribution of μ is normal, with variance

$$S(t) \equiv \left(\frac{1}{\sigma_\mu^2} + \frac{1}{\sigma_m^2} + \frac{t}{\sigma_x^2}\right)^{-1} \tag{1}$$

and with mean

$$E(\mu|m, x, t) \equiv S(t)\left(\frac{\bar{\mu}}{\sigma_\mu^2} + \frac{m}{\sigma_m^2} + \frac{x}{\sigma_x^2}\right). \tag{2}$$

The parameters σ_μ^2, σ_m^2, σ_x^2, and $\bar{\mu}$ are known. The sufficient statistics are m, x, and t.

Search On and Off the Job

In my early work on matching (Jovanovic 1979*a*), the worker never chose to be unemployed—he could start a new job whenever he wished. Here, we assume instead that his contacts with employers occur as a Poisson process, with rate λ if he works and rate δ if he does not. Both λ and δ are exogenous and costless; if the worker finds it easier to search when unemployed, $\lambda < \delta$. Further, he has *no recall* to employers by whom he was screened previously or to employers he has worked for before.

The quality of a new match is independent of the quality of the old one. When a contact is made, the employer screens the worker. On seeing the outcome of the screen, the worker decides right away either to take the new job or to return to his old one.

Is this search technology reasonable? One could raise several objections. First, search is neither costless nor exogenous.[1] Burdett (1978), for example, assumes that it is cheaper to search when unemployed. Moreover, he allows workers the option of not searching at all. Our assumption of differential efficiency of on-the-job and off-the-job search is designed to reflect the behavioral effect of differential search costs in the two states. But if the intensity of search is significantly related to the perceived quality of the match, the job-to-job transition rate will be misspecified. We shall return to this point in the next section.

Second, this model has it that, apart from unemployment compensation, the only reason why the worker would choose unemployment is that the chance of getting a job offer is higher ($\delta > \lambda$). Is this assumption justified? An unemployed worker does have more time to search, but he is also "further removed" from the marketplace so that the time spent searching may not be as productive. Flinn and Heckman (1983) find evidence that, for young men, $\delta > \lambda$, although their conclusions are conditional on the truth of a simpler search model.

Third, the worker would like the option to recall past offers. Sup-

---

[1] In a model with costly and endogenous on-the-job search (Jovanovic 1979*b*) it is shown that the well-matched workers and workers with relatively more firm-specific human capital spend less time searching

pose he quits job A because a better offer arrives for job B. If job B turns out to be a disappointment, he will want to return to job A unless another decent offer has turned up in the meantime. Moreover, the no-recall assumption rules out temporary layoffs, which are so common in U.S. manufacturing. With recall, some temporary separations *would* take place in this model. Nevertheless, only a small fraction of the temporary separations one observes can, I think, be explained by a model in which product demand does not fluctuate. A separation caused by a mismatch is likely to be permanent.

PREFERENCES: Workers and firms are risk neutral and infinitely lived. They discount the future at a common, constant discount rate.

EQUILIBRIUM: Firms bid for workers by offering them lifetime contracts that, by assumption, are honored ex post. The same contracting assumption is used in Jovanovic (1979*a*), where it is shown that, since this mechanism transfers all the rents from the match to the worker, one equilibrium is for the firm always to pay the worker his expected marginal product, so that

$$w = E(\mu | m, x, t) \tag{3}$$

is the worker's pay at each time. This wage contract (i.e., a function that relates sufficient statistics to wage payments) is, however, not unique. Only the turnover behavior (i.e., the function that relates sufficient statistics to quit decisions) is unique (see Jovanovic 1979*a*, theorem 2). I shall refer to *w* as "the wage," even though wages and marginal products may diverge.

A troubling feature of contracts that bestow all rents to t
is that they are not *consistent* ("self-enforcing")—the firm w
cheat and pay well-matched workers less than their produc
enough to keep them. One standard reply to this is that
cheats one worker will develop a bad reputation in the
large, as well as with the rest of its own work force. A second w
prevent cheating is to make contracts (such as the one in eq. (3))
explicit and legally enforceable. In the absence of such constraints on firm behavior, the only *perfect* equilibrium in this model is one in which wages are driven down to the workers' value of leisure, and no turnover and unemployment take place! I shall assume that contracts are legally enforceable or, equivalently, that the firm's incentive to cheat is outweighed by its concern for its reputation.

The martingale property of conditional expectations implies that the worker's current job wage is a random process without drift. In fact, let

$$\pi(t) = \frac{1}{(1/\sigma_\mu^2) + (1/\sigma_m^2)} - \frac{1}{(1/\sigma_\mu^2) + (1/\sigma_m^2) + (t/\sigma_x^2)}.$$

This is the posterior variance, subtracted from the prior (but post-screen) variance. It increases monotonically from zero to $(\sigma_\mu^{-2} + \sigma_m^{-2})^{-1}$. Chernoff (1968, lemma 4.1) shows that the wage is a standard Wiener process in the $\pi(t)$ scale. That is, $w = Z[\pi(t)] + E(\mu|m, 0, 0)$, where, of course, $Z(0) = 0$.

The three sufficient statistics (m, x, t) can be reduced to just two: the posterior mean and posterior variance given in equations (1) and (2) or, equivalently, equations (1) and (3). Since $S(t)$ is strictly decreasing in t, one can use t in place of $S(t)$ for one of the sufficient statistics, so that our pair of sufficient statistics is (w, t)—the worker's current wage and his current length of job tenure. In this stationary world, neither his age nor anything else in the worker's employment history will affect his behavior.

While unemployed, the worker gets an unemployment insurance benefit each period, for as long as he stays unemployed. Let u be the combined value of leisure and the unemployment benefit per period. The present value of becoming unemployed is constant over time and is denoted by U.

When a new job offer arrives, it will have an initial wage equal to $E(\mu|m', 0, 0)$, where m' is the outcome of the screen for the match between the worker and that particular job. Since both $\bar{\mu}$ and m are normally distributed, the initial wage is also normally distributed with mean $\bar{\mu}$ and variance $\sigma_\mu^4/(\sigma_\mu^2 + \sigma_m^2)$. Let $F(w)$ denote this distribution. Search on the job as well as search off the job yield offers that are drawn from F.

Let $V(w, t)$ be the present value of income when the worker is on a match with sufficient statistics (w, t). If the cost of job change is zero, the worker will change jobs if his on-the-job search yields him an *initial* wage offer w' such that $V(w', 0) > V(w, t)$. If the reverse inequality holds, he will reject the offer and continue working on his current job. This defines a reservation wage $\phi(w, t)$ that solves

$$V(w, t) = V[\phi(w, t), 0]. \tag{4}$$

This is a reservation wage in the usual sense: if $w' \leq \phi$, the outside offer is rejected. However, while the worker's current job situation (summarized by w and t) may change over time, the worker's outside opportunities (summarized by λ and F) remain unchanged. The job-to-job hazard rate is

$$\hat{\lambda}(w, t) \equiv \lambda\{1 - F[\phi(w, t)]\}. \tag{5}$$

If the current match is a disappointment and if acceptable new offers do not arrive, the worker may choose unemployment. Let $\theta(t)$ be the lowest wage the worker will accept on his current job and still remain there. At tenure t, a wage lower than $\theta(t)$ will lead the worker to quit and become unemployed. Then $\theta(t)$ solves the equation

$$V[\theta(t), t] = U. \tag{6}$$

Finally, let ψ be the smallest wage offer that will induce the worker to leave unemployment and accept the job. Then ψ solves

$$U = V(\psi, 0). \tag{7}$$

The hazard rate from unemployment to employment is just $\delta[1 - F(\psi)]$. The present value of unemployment satisfies the equation

$$U = \int_0^\infty \left\{\int_0^T e^{-rt} u\,dt + e^{-rT}\left[\int_\psi^\infty V(x, 0) f(x)dx + F(\psi)U\right]\right\} \delta e^{-\delta T} dT,$$

which simplifies to

$$U = \frac{1}{r + \delta[1 - F(\psi)]}\left[u + \delta\int_\psi^\infty V(x, 0) f(x)dx\right]. \tag{8}$$

If $\delta = 0$ (no offers ever arrive), $U = u/r$.

The corresponding equation for V requires some detailed development and is given in equation (A1), which defines V in terms of a parametrically given value of U. Equations (8) and (A1) provide us with two equations that are to be solved for V and U. In the Appendix to the paper it is shown that these equations have a unique solution for V and U if

$$\frac{\lambda}{r}\left(1 + \frac{\sigma_\mu^1}{\sigma_\mu^2 + \sigma_m^2}\right) < 1. \tag{9}$$

The second term in parentheses is the variance of the w distribution, and λ is the speed with which offers from that tion come in. Equation (9) says that search opportunities mu too favorable compared to the rate of discount, r. One sh that this condition is sufficient but not necessary.[2]

For $\tau \geq t$, let $H(w'; \tau, w, t) = \text{prob}\,(w_\tau \leq w' | w_t = w)$.[3] This is th probability that, given (w, t) today, the worker will still be in his current firm at time τ, and that at that time his wages will not exceed w'.

---

[2] One could weaken condition (9) considerably by assuming that, instead of $X(t)$, cumulative output is given by $\int_0^t Q[dX(s)]$, when Q is some bounded function of the instantaneous mutually observed latent variable increment. This does not alter the learning structure, and it puts an upper bound on the instantaneous reward in the value function, thereby making existence proofs easier. On the other hand, it also introduces an arbitrary transformation Q that complicates the rest of the analysis, and I shall abstain from using it here.

[3] This probability is rigorously derived in the Appendix.

Let $p(\tau, w, t) \equiv H(\infty; \tau, w, t)$. Then p is the probability that, given (w, t) today, the worker will not have left by time τ.

An increase in the current wage w increases p for two reasons. First, the higher the current wage, the less likely it is that at any future time it will fall to the point $\theta(\cdot)$ at which the worker will choose unemployment. Second, the higher the wage is today, the higher it is likely to be in the future, and the less likely it is that the worker will be bid away by another firm. In other words, both types of transition are more likely for the lower-paid workers.

The value function is differentiable in w and t, with $(\partial V/\partial w)(w, t) = \int_t^\infty e^{-r(\tau - t)} p(\tau, w, t) d\tau$, which is positive for $w > \theta(t)$. Since the wage is a random walk, a unit increase today raises expected wages by a unit in every future period. The unit increase is, however, enjoyed only for as long as the worker stays with the firm. Since p increases with w, so does $\partial V/\partial w$. In other words, V is *convex* in w.

Since the wage process can be written as $w = Z[\pi(t)] + E(\mu|m, 0, 0)$, one can write wages in differential form:[1] $dw(t) = \sqrt{d\pi/dt}\, dZ[\pi(t)] = [S(t)/\sigma] dZ[\pi(t)]$. The standard deviation of the wage increments, $S(t)/\sigma$, declines with t. Now V depends on t only through the effect of t on $S(t)$. Since V is convex in the wage, we conclude that $\partial V/\partial t < 0$. The risk-neutral worker faces a wage that follows a driftless process. He is, however, insured against the downside risk by the option of becoming unemployed and by the possibility of getting a wage offer from another employer.

III. Reservation Wages and Hazard Rates

The three reservation wages are $\phi(w, t)$, $\theta(t)$, and ψ. They pertain, respectively, to job-to-job, job-to-unemployment, and unemployment-to-job transitions.

Job to Job

Since $\partial V/\partial w > 0$ and $\partial V/\partial t < 0$, differentiation in equation (4) with respect to w and to t yields

$$\frac{\partial \phi}{\partial w} > 0, \quad \frac{\partial \phi}{\partial t} < 0, \quad \phi(w, 0) = w. \tag{10}$$

This implies that wages and reservation wages diverge from each other as t increases; $\phi(w, t) < w$ for $t > 0$. This means that after he has worked somewhere for a while, the worker will be willing to accept a job with a lower current wage. This is a prediction that was also present in human capital models in which a worker would trade a job with a stable wage for a job on which his initial wage would be lower, but because of on-the-job training the new wage would grow and "overtake" the old wage.[5] Our prediction has the same flavor: the worker would give up a job with a stable wage in favor of a job with a lower initial wage, so long as the new job offered the *possibility* of growth. Here, job tenure t is an index of future wage stability; as t increases, the worker becomes less demanding elsewhere—hence the decline in ϕ. Evidence indicates that quitting to take a job with an initially lower wage is quite widespread, especially for older men (Bartel and Borjas 1981).

---

[1] See Cox and Miller (1965, p. 228) for details on transformations of the Wiener process.

The job-to-job hazard rate is given in equation (5). It decreases with w and increases in t: $\partial\hat{\lambda}/\partial w = -\lambda f(\partial\phi/\partial w) < 0$ and $\partial\hat{\lambda}/\partial t = -\lambda f(\partial\phi/\partial t) > 0$. When conditioned on the wage, $\hat{\lambda}$ exhibits *positive* duration dependence. This prediction too has been made, less formally, by the human capital theorists. Their argument assumes that general skills, transferable across firms, are accumulated on the job. If the worker's wage is held constant while the worker's general productivity grows, his alternative wage will rise and he will be more likely to leave.[6]

Equations (4), (6), and (7) imply that $\phi[\theta(t), t] = \psi$. When $w = \theta(t)$, the worker is indifferent between staying on and becoming unemployed. But he is also indifferent between staying unemployed and taking a job with an initial wage of ψ. Since $\theta(t)$ is a lower bound on w, ψ is a lower bound on $\phi(w, t)$.

Unemployment to Job

The unemployed worker's reservation wage is w. The rate

ment from unemployment to employment is $\delta[1 - F(\psi)]$. S

$t) \geq \psi$, this rate exceeds the job-to-job rate if $\lambda \leq \delta$.

The constancy of the unemployment hazard reflects the assu

permanence of unemployment insurance. It is, of course, desirable that theory should reflect the temporary duration of unemployment benefits, but my attempts to take account of it quickly led to an unexpected complexity in the analysis. The reason is that, in addition to endowing the unemployed worker with temporary unemployment benefits, one must also insist on a minimum duration of *continuous* employment prior to the unemployment spell. Without this requirement, as the expiration date of his benefit approached, the unemployed worker could accept *any* job for an *instant* and then reenter unemployment immediately, thereby assuring himself of another maximum period of coverage.

---

[5] See Mincer (1974), esp. the references to overtaking

[6] See, e g., Becker (1962) To get this effect in the human capital model, however, one must also explain why the worker's wage is being held constant while his general productivity is increasing.

Job to Unemployment

Since $\partial V/\partial t < 0$ for $w > \theta(t)$, equation (6) implies that $\theta(t)$ is increasing in t. *Equations (6) and (7) imply that* $\theta(0) = \psi$. *Therefore* ψ *is a lower* bound for both ϕ and θ. In other words, the worker is least demanding when he is unemployed. This is not surprising—unemployment is something the worker can always have.

The job-to-unemployment hazard is hard to characterize. The only result is that it must initially increase.[7] Flinn and Heckman (1983) find that when one controls for unobserved heterogeneity, the job-to-unemployment hazard does indeed show an initial increase, but that it eventually declines. This finding agrees exactly with the prediction in Jovanovic (1979*a*). That prediction was based, however, on there being a positive probability that the worker would remain on his current job forever. In the present model, every match ends with probability one: even if unemployment were never to become an attractive option, the worker is sure of eventually getting a better offer from another employer.

IV. Learning by Inspection versus Learning from Experience

Recall that if $\sigma_m = 0$ all learning about the match is entirely by inspection (i.e., through the screen), whereas if $\sigma_m = \infty$, all learning is from experience.

Perfect Screen ($\sigma_m = 0$)

When all learning is by inspection, the wage on a job is constant and V does not depend on t. None of the hazards exhibit any duration dependence; in fact, since V (on a given match) and U are both constant, the job-to-unemployment hazard is zero. This version of the model is similar to Burdett (1978) and, in some respects, to Flinn and Heckman (1982).

---

[7] For any distribution $G(t)$ with density $g(t)$ $(t \geq 0)$ with $g(0) = 0$, the hazard $g(t)/[1 - G(t)]$ increases at $t = 0$ because g does. In the present context g is a transform of the first passage density of the process defined in eq. (3) through the boundary $\theta(t)$. First passage densities for diffusion processes have the property that $g(0) = 0$ so long as $w(0) > \theta(0)$.

No Screen ($\sigma_m = \infty$)

In this case, the only way to learn is from on-the-job experience. The variance of the initial wage-offer distribution (F) is $\sigma_\mu^1/(\sigma_\mu^2 + \sigma_m^2)$, and it tends to zero as σ_m gets large. A striking difference between this polar case and the previous one is in the behavior of wages: in the experience case, all wage growth takes place on the job, while in the inspection case all wage growth takes place between jobs.

In the general case, this model gives us a theory of on-the-job and off-the-job wage growth.

V. The Semi-Markov Property

As a two-state model of employment and unemployment, this model is semi-Markov: only the state and the length of time spent in that state matter. This is trivially true for the unemployed state in which even duration does not affect the probability of leaving the state. The transition to employment can take place only from unemployment, and the initial conditions are fully specified by the initial wage draw. The distribution of the initial wage draw is $f(w)/[1 - F(\psi)]$; since ψ is a constant (independent of history), this distribution does not depend on the worker's history. In the employed state as well, neither the initial conditions nor the subsequent evolution is affected by the history. Thus the model is semi-Markov.

Job durations, on the other hand, are distributed neither identically nor independently, except if $\sigma_m = \infty$. When $\sigma_m = \infty$, $F(w)$ has zero variance; all jobs appear identical ex ante, and match quality is independent from one draw to the next.[8] When $\sigma_m < \infty$, this is n
true. To see why not, assume that $\sigma_m = 0$. Conditional on t
the distribution of job durations is exponential, with mea
$F(w)]\}^{-1}$ (recall that in this special case neither w nor ϕ depe
Since successive jobs start at successively higher wages, the se
is likely to last longer than the first, even when one does *not* c
on wages. This shows why job durations are not *identically* distri
To see why independence does not hold, assume that the first job lasted for an unusually long time. This makes it likely that the wage on the first job was higher than the average first-job wage. The wage on the second job must have exceeded the first-job wage, and since the latter was unusually high, the second-job wage is also likely to be higher than usual. Therefore, the duration of the second job is likely

---

[8] When $\sigma_m = \infty$, job durations follow a pure renewal process. See Jovanovic (1979*a*, p. 982) for further details.

to be longer than usual. And so the durations of the first and second jobs are positively correlated, so long as one does not condition on wages.

This correlation will be preserved for the general case when $\hat{\sigma} > 0$, which means that movements from job to job do not possess the Markov property. As a two-state model, however, the model is semi-Markov.[9]

Appendix

We begin with a representation of the wage process on a given job. This process is defined by equation (3) and by the structure generating the sufficient statistics (m, x, t). The process is defined independently of the worker's actions.

Let $(\Omega, \mathcal{F}, P)$ be a probability space, with $\omega \in \Omega$ the elementary events. $\mathcal{F}$ is a sigma algebra of subsets of Ω; P assigns probability to sets in $\mathcal{F}$. For $t \in [0, \infty)$ let $w(t, \omega)$ be the sample function of the wage given that ω occurs. Since the wage is a Wiener process, the sample functions are continuous with probability one (Feller 1971, p. 181). For $\tau \geqslant t > 0$, $w \in R$, and any continuous function $\theta(t)$, let $A(B', \tau, B, t; \theta) \equiv \{\omega : w(t, \omega) \in B, w(\tau, \omega) \in B', w(s, \omega) \geqslant \theta(s)$ for $s \in [t, \tau]\}$, where B' and B are Borel subsets of the real line. The sample functions $w(t, \omega)$ are defined independently of whether the worker stays or leaves. Let

$$r(\tau, t, \omega, \phi) = 1 - \exp\left\{\lambda \int_t^\tau \left(1 - F\{\phi[w(s, \omega), s]\}\right) ds\right\}.$$

This is the probability that the sample path $w(t, \omega)$ will be "extinguished" on $[t, \tau]$ because a better job offer has been found. Therefore, of those paths that start in B and end up in B', the fraction extinguished is $R(B', \tau, B, t, \phi, \theta) = \int_A r(\tau, t, \omega, \phi) P(d\omega)$. Let $H(B'; \tau, w, t, \theta, \phi)$ be the probability that the wage function (which starts from w at t) does not get extinguished (either because the worker takes a better job or because he chooses unemployment) by time τ *and* that $w(\tau, \omega) \in B'$. Then for any Borel subset B of the real line, H satisfies $(1 - R)P(A) = \int_B H(B'; \tau, w, t, \theta, \phi) D(dw; t)$, when $D(w; t) \equiv P\{\omega : w(t, \omega) < w\}$ is the unconditional probability of wages at t. The solution for H exists and is unique (Feller 1981, p. 160).

After this rather mechanical development, we are ready for the worker's decision problem when employed. While in his current firm, his instantaneous payoff is the wage. If he leaves to become unemployed, his payoff is U in present value. If he leaves for a better job, he gets $V(w', 0)$ in present value, where w' is the starting wage on the next job. The measure of exit to other jobs at τ $(\tau > t)$ is $\int_{\theta(\tau)}^{\infty} \lambda\{1 - F[\phi(s, \tau)]\} H(ds; \tau, w, t, \theta, \phi)$. The measure of exit to unemployment is as yet undefined. To avoid further algebra, we shall employ the following "trick": Let Π_1 and Π_2 be two payoffs, and p_1 and p_2 the respective probabilities of those payoffs. Then $p_1\Pi_1 + p_2\Pi_2 = (p_1 + p_2)\Pi_1 + p_2(\Pi_2 - \Pi_1)$ holds *identically*. Now interpret Π_1 as the present value of unemployment and Π_2 as the present value of getting a better job. We shall say that if the current job is terminated (which happens with probability $p_1 + p_2$), the worker gets U. But if termination is because of a better job, he gets $V(w', 0) - U$. In view of the above-mentioned identity, this reformulation alters neither the worker's optimal policy nor his payoff. Thus,

---

[9] Heckman and Borjas (1980) present evidence against the semi-Markov model by finding what they call "occurrence" dependence with unemployment spells: the pattern of *past* unemployment experience seems to affect how long the *present* unemployment spell will last.

$$V(w, t) = \max_{\theta, \phi}\left[\int_t^\infty e^{-r(\tau - t)}\xi(\tau, w, t)d\tau + U\int_t^\infty e^{-r(\tau - t)}\bar{H}(d\tau, w, t, \theta, \phi)\right], \tag{A1}$$

where

$$\xi(\tau; w, t) = \int_{\theta(\tau)}^\infty \left\{x + \lambda\int_{\phi(\tau, x)}^\infty [V(w', 0) - U]f(w')dw'\right\}H(dx; \tau, w, t, \theta, \phi),$$

where $\bar{H}(\tau, \ldots) = 1 - \int_{\theta(\tau)}^\infty H(dw'; \tau, \ldots)$ is the probability that exit (for either reason) takes place on $[t, \tau]$, and where θ and ϕ are restricted to be continuous. Note that $H = 0$ and $\bar{H} = 1$ everywhere if $w < \theta(t)$. The standard contraction mapping theorems (Blackwell 1965) cannot be applied to the operator on the right-hand side of equation (A1) because ξ is not bounded in w. Dividing both sides of equation (A1) by $\epsilon + (w - \bar{\mu})^2$ (for some $\epsilon > 0$) and defining $\tilde{V} = V/[\epsilon + (w - \bar{\mu})^2]$ yields

$$\tilde{V}(w, t) = \max_{\theta\, \phi}\left[\int_t^\infty e^{-r(\tau - t)}\tilde{\xi}d\tau + \frac{U}{\epsilon + (w - \bar{\mu})^2}\int_t^\infty e^{-r(\tau - t)}\bar{H}\right], \tag{A2}$$

where

$$\tilde{\xi} = \frac{\xi}{\epsilon + (w - \bar{\mu})^2}$$

$$= \int_{\theta(\tau)}^\infty \left\{\frac{x}{\epsilon + (w - \bar{\mu})^2} + \lambda\int_\phi^\infty \left[\frac{\epsilon + (w' - \bar{\mu})^2}{\epsilon + (w - \bar{\mu})^2}\tilde{V}(w', 0) - U\right]f(w')dw'\right\}$$

To prove existence of $\tilde{V}$ one must show that the operator on the right-hand side of equation (A2) is (i) continuous in (w, t), (ii) bounded in (w, t), and (iii) a contraction mapping.

i) H is continuous because it is the distribution of a diffusion process truncated by continuous boundaries (Cox and Miller 1965). Evidently, $\tilde{\xi}$ is continuous in both w and t.

ii) $\tilde{\xi}$ certainly exists for any bounded $\tilde{V}$; f is normal, and H is essentially a "scaled-down" normal distribution. Moreover, for any constant a, $\int_a^\infty xH(dx, \tau, w, \ldots)$ is $O(w)$ because the variance of the wage process is unaffected by its position. In fact, $\tilde{\xi} \to 0$ as $|w| \to \infty$.

iii) Let $V_1, V_2 : R^2 \to R^1$ be two bounded and continuous functions. Let $\tilde{\xi}_1$

and $\tilde{\xi}_2$ denote the two resulting values of $\tilde{\xi}$ when evaluated at V_1 and V_2, respectively. Let $\|\cdot\| \equiv \sup |\cdot|$ $(w \in R^1, t \in R^1_+)$. Then

$$\|\tilde{\xi}_1 - \tilde{\xi}_2\| \leq \lambda \sup_{w,t} \int_{\phi(w,t)}^{\infty} \frac{\epsilon + (w' - \bar{\mu})^2}{\epsilon + (w - \bar{\mu})^2} |V_1(w', 0) - V_2(w', 0)| f(w')dw'$$

$$\leq \lambda \sup_{w,t} \int_{-\infty}^{\infty} \frac{\epsilon + (w' - \bar{\mu})^2}{\epsilon + (w - \bar{\mu})^2} |V_1(w', 0) - V_2(w', 0)| f dw'$$

$$\leq \lambda \left[1 + \frac{\sigma_\mu^4}{(\sigma_\mu^2 + \sigma_m^2)}\right] \sup_{w'} |V_1(w', 0) - V_2(w', 0)|$$

when $\epsilon = 1$. The last inequality follows because $\int[1 + (w')^2] f(w')dw$ is equal to the term in square brackets in the expression above. We have proved that $\|\tilde{\xi}_1 - \tilde{\xi}_2\| \leq k\|V_1 - V_2\|$ if $\lambda\{1 + [\sigma_\mu^4/(\sigma_\mu^2 + \sigma_m^2)]\} \equiv k$. Then equation (A2) implies that the contraction property holds if $k < r$. This is equivalent to the condition in equation (9). This condition is sufficient; by varying ϵ away from 1, weaker (and more complicated) conditions can be derived.

Having proved the existence of a unique $\tilde{V}(w, t)$, we then obtain a unique $V(w, t) = [\epsilon + (w - \bar{\mu})^2]\tilde{V}(w, t)$ satisfying equation (A1).

We have, however, a fixed-point problem in a function $V(w, t)$ and a scalar U. The solution for V obtained from equation (A1) depends parametrically on U. Let us make this explicit by writing it as $V(U)$. We shall now prove that

$$\|V(U_1) - V(U_2)\| \leq |U_1 - U_2|. \tag{A3}$$

An application of the envelope theorem shows that V is differentiable with respect to U and that $(\partial V/\partial U)(w, t) = \int_t^\infty e^{-r(\tau - t)}\hat{p}(d\tau; w, t)$, where $\hat{p}(\tau; w, t)$ is the probability that a transition to unemployment has already taken place by time τ, given (w, t) today. The proof is parallel to that of $\partial V/\partial w$, and so it is omitted here. Now for all (w, t), $\int_t^\infty \hat{p}(d\tau; w, t) \leq 1$, and therefore $\partial V/\partial U \leq 1$ for all (w, t).[10] Since the result also holds for all U, equation (A3) follows.

The result in equation (A3) shows that the operator defined on the right-hand side of equation (8) is a contraction mapping (on scalars U) with modulus $\delta[1 - F(\psi)]/\{r + \delta[1 - F(\psi)]\} < 1$. This completes our proof of the existence of equilibrium.

We turn now to differentiability of V. The proof of differentiability with respect to w heavily utilizes the fact that the evolution of the wage increments does not depend on its initial position, w. Let $\nu(\tau, \omega) = w(\tau, \omega) - w(t, \omega)$. Then, since the increments of w do not depend on its position, $H(w'; \tau, w, t, \theta, \phi) = H(\nu; \tau, 0, t, \theta - w, \phi'')$, where $\phi'' = \phi(w + \nu, s) - w$. If the current wage is $w + \rho$, the stopping boundaries $\theta(s) + \rho$ and $\phi(x - \rho, t) + \rho$ are feasible. I have chosen these hypothetical feasible boundaries so as to keep $H(\nu, \ldots)$ unchanged. Since the hypothetical boundaries are not necessarily optimal,

---

[10] One can also show that $p(\tau; w, t)$ is increasing in U and that V is therefore *convex* in U. The intuition is simple. When U is very small, the option of unemployment is not likely to be used, and a one dollar increase in U will not raise the employed worker's welfare by much. But an increase in U raises $\theta(t)$ uniformly (see eq. [6]) so that when U is larger and θ is closer to w, a dollar increase in U is worth more in present value.

$V(w + \rho, t)$ exceeds the expression on the right-hand side of equation (A1) when it is evaluated at these boundaries. Therefore,

$$V(w + \rho, t) - V(w, t) \geqslant \rho \int_t^\infty e^{-r(\tau - t)} \int_{\theta(\tau)-w}^\infty H(d\nu; \tau, 0, t, \phi^w, \theta - w)d\tau.$$

A similar argument using policies optimal when the wage is $w + \rho$ leads to

$$V(w + \rho, t) - V(w, t) \leqslant \rho \int_t^\infty e^{-r(\tau - t)} \int_{\theta(\tau)-w-\rho}^\infty H(d\nu; \tau, 0, t, \phi^{w+\rho}, \theta - w - \rho)d\tau.$$

Dividing both expressions by ρ and taking the limit as $\rho \rightarrow 0$ gives us existence of the derivative:

$$\frac{\partial V}{\partial w}(w, t) = \int_t^\infty e^{-r(\tau - t)} \int_{\theta(\tau)-w}^\infty H(d\nu; \tau, 0, t, \phi^w, \theta - w). \qquad \text{(A4)}$$

This expression is strictly positive whenever the wage exceeds $\theta(t)$. As $w \rightarrow \theta(t)$, however, $H(\nu, \tau; \ldots) \rightarrow 0$ for all $\tau > t$, so that $(\partial V/\partial w)[\theta(t), t] = 0$—the well-known condition of "smooth fit" in optimal stopping problems, as it applies in our case

Rewriting $\partial V/\partial w$ in terms of the cumulative wage distribution,

$$\frac{\partial V}{\partial w}(w, t) = \int_t^\infty e^{-r(\tau - t)} \int_{\theta(\tau)}^\infty H(dw; \tau, w, t, \phi, \theta).$$

The proof of the *existence* of $\partial V/\partial t$ goes the same way; the key again is to transform the problem in such a way that neither the evolution of the wage process nor the feasible set for the controls (θ, ϕ) is affected by t This can be done by using the standard Wiener property of the wage when the time scale is changed from $[t, \infty)$ to $[0, S(t))$ Since the same procedure is used in Jovanovic (1979*a*, app.), we shall not repeat it here. The derivative that one obtains using this procedure cannot, however, be signed for this mor plicated problem.

References

Bartel, Ann P., and Borjas, George J. "Wage Growth and Job Tur Empirical Analysis." In *Studies in Labor Markets*, edited by Sherw Chicago: Univ. Chicago Press (for Nat. Bur. Econ. Res.), 1981

Becker, Gary S "Investment in Human Capital: A Theoretical Analysis *J.P.E.* 70, no. 5, pt. 2 (October 1962): 9–49.

Blackwell, David. "Discounted Dynamic Programming." *Ann. Math. Statis* 36 (February 1965): 226–35.

Burdett, Kenneth. "A Theory of Employee Job Search and Quit Rates" *A.E.R.* 68 (March 1978): 212–20.

Chernoff, Herman. "Optimal Stochastic Control." *Sankhya*, ser. A, 13 (June 1968): 111–12.

Cox, David R., and Miller, Hilton D. *The Theory of Stochastic Processes* New York: Wiley, 1965.

Feller, William. *An Introduction to Probability Theory and Its Applications*. Vol. 2. 2d ed. New York: Wiley, 1971.

Flinn, Christopher J.; and Heckman, James J. "New Methods for Analyzing

Structural Models of Labor Force Dynamics." *J. Econometrics* 18 (January 1982): 115–68.

———. "Are Unemployment and Out of the Labor Force Behaviorally Distinct Labor Force States?" *J. Labor Econ.* 1 (January 1983): 28–42.

Heckman, James J., and Borjas, George J. "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence." *Economica* 47 (August 1980): 247–83.

Jovanovic, Boyan. "Job Matching and the Theory of Turnover." *J.P.E.* 87, no. 5, pt. 1 (October 1979): 972–90. (*a*)

———. "Firm-specific Capital and Turnover." *J.P.E.* 87 (December 1979): 1246–60. (*b*)

Mincer, Jacob A. *Schooling, Experience, and Earnings.* New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1974.

Toikka, Richard S. "A Markovian Model of Labor Market Decisions by Workers." *A.E.R.* 66 (December 1976): 821–34.

Are Bond-financed Deficits Inflationary? A Ricardian Analysis

Bennett T. McCallum

Carnegie-Mellon University and National Bureau of Economic Research

This paper considers the possible theoretical validity of the following "monetarist hypothesis": that a constant, positive government budget deficit can be maintained permanently and without inflation if it is financed by the issue of bonds rather than money. The question is studied in a discrete-time, perfect-foresight version of the competitive equilibrium model of Sidrauski, modified by the inclusion of government bonds as a third asset. It is shown that the monetarist hypothesis is invalid if the deficit is defined exclusive of interest payments, but that it is valid under the conventional definition. It is also shown that the stock of bonds can grow indefinitely at a rate in excess of the rate of output growth, pro that the difference is less than the rate of time preference formal conclusions do not take account of possible limitatio taxing capacity arising from default incentives that tend to gro the stock of bonds.

I. Introduction

Recent developments in U.S. monetary and fiscal affairs have led to a renewal of interest in issues reminiscent of the "monetarist vs. Keynesian" debates of previous decades. In particular, the Council of Economic Advisers' 1982 and 1983 forecasts of a long string of unusually

This paper was prepared for the Second Peterkin Symposium, "Foundations of Monetary Policy and Government Finance," held April 1–2, 1982, at Rice University. I am indebted to Robert Barro, John Bryant, Marvin Goodfriend, Peter Neary, Edmund Phelps, Paul Romer, José Scheinkman, Warren Weber, and two anonymous referees for helpful comments on earlier versions. Financial support was provided by the National Science Foundation (SES 82-08151)

[*Journal of Political Economy*, 1984, vol. 92, no. 1]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0012$01.50

large federal budget deficits, together with the Federal Reserve's repeated avowals to keep monetary growth rates low, have conferred intense practical interest upon the question of whether bond-financed deficits have significant impact on aggregate demand and thereby on price level and/or output magnitudes. In addition, recent academic emphasis on improved choice-theoretic foundations for aggregate specifications has led to some new notions regarding monetarist models, most prominently in the work of Sargent and Wallace (1981, 1982).

The present paper seeks to contribute to the theoretical understanding of such issues by consideration of a precisely specified question that bears directly on the monetarist notion that bond-financed deficits are unimportant for aggregate demand. In particular, it asks whether it is possible that—in an equilibrium model of a nongrowing economy with maximizing agents and cleared markets—a positive government budget deficit can be maintained permanently and without inflation if financed entirely by bond sales, the money stock being held constant. As it transpires, the answer to this question depends on the way "deficit" is defined, the crucial distinction involving the inclusion or exclusion of interest payments on outstanding government debt.

II. Preliminaries

In order to limit the issues at hand in a manageable fashion, it will be presumed throughout the analysis that the economy can be represented by a deterministic, aggregative, flexible-price, equilibrium model. For some issues, such a model might be inadequate or misleading. Our present concern, however, involves the influence on inflation rates of a policy stance maintained over an extended period of time. For that type of concern, a flexible-price equilibrium model—which presumes that aggregate demand effects are manifested primarily in price level or inflation responses—seems well suited.

As a matter of terminology—and to sharpen the issues—let us define a *monetarist viewpoint* as one that asserts that bond-financed deficits have no effect on aggregate demand. More precisely, our *monetarist hypothesis* is that, for given time paths of the money stock and government spending, it does not matter for aggregate demand whether the necessary revenue is raised by taxation or by bond sales. In other words, bond-financed changes in tax receipts have—*according to the monetarist hypothesis—no effects on the price level* or on output.

At this point, it perhaps needs to be asked whether there is *any* explicable reason to believe that the monetarist hypothesis (as defined) might be correct. Discussions in the most well-known references notwithstanding,[1] the main intellectual support for such a position seems to be provided by the "Ricardian equivalence theorem," which obtains in some models in which infinite-lived agents correctly take account of the effects on future budgets of current budgetary actions.[2] With fixed time paths of government spending and money creation, any bond-financed change in current taxes implies changes in future interest payments to be made by the government. If these are to be financed by lump-sum taxes and if the government and private agents face the same market interest rates, then the change will have no effect on a representative private agent's intertemporal budget constraint. Under such conditions, then, a bond-financed change in taxes will have no effect on the agent's supplies or demands and, consequently, no effect on the price level or output—just as predicted by the monetarist hypothesis.

The foregoing statement of the Ricardian result presumes that agents have infinite life spans, which is obviously untrue. But, as is well known, Barro (1974) has demonstrated that an economy of finite-lived agents who care about the utility of their offspring or parents may, under reasonably general conditions, be treated for analytical purposes as one with infinite-lived agents.[3] Consequently, this feature of the analysis seems acceptable, given the aims of the investigation.

A second crucial assumption of the Ricardian analysis is that agents are cognizant of effects on their own intertemporal budget constraints of governmental debt issues. But this assumption is merely a particular application of the hypothesis of rational expectations, the merits of which have been detailed extensively elsewhere.

Other complicating aspects of reality—uncertainty, distrib fects, multiple interest rates—are also ignored in the R equivalence argument. But the same is true of most policy- theoretical analyses of macroeconomic phenomena. As the apparent reason why the issue at hand requires a different treatment, it would seem satisfactory to neglect them here, as elsewhere.[4]

---

[1] Many of these are included in Gordon (1974) and Stein (1976). The models typically assume that consumption or private expenditure depends directly upon private wealth, with government debt included as a component of the latter (see, e.g., Brunner and Meltzer 1976, p. 72). In these models, then, a bond-financed tax reduction directly increases aggregate demand.

[2] In this analysis, and in what follows here, taxes are assumed to be of the lump-sum variety.

[3] This statement presumes that intergenerational transfers are operative. A discussion of circumstances under which Barro's result is inapplicable is provided by Drazen (1978).

[4] Barro (1974, 1983) argues that neglect of these complicating features does not serve to distort the results in a predictable direction.

The discussion thus far seems to suggest that bond-financed deficits could indeed be noninflationary. Each bond issue/tax reduction package has no impact on aggregate demand or the price level, so a sequence of bond-financed tax reductions should apparently have no impact on the inflation rate. But it is possible that the situation is different in the case of a *permanent* deficit, financed by indefinitely continuing issuance of bonds. In that case, as the deficit continues the outstanding bond stock continues to grow. So, accordingly, does the interest that must be paid each period on the outstanding stock. If, for example, the magnitude of the real deficit (net of interest payments) is kept at d, the real bond stock b_t will be required to grow according to

$$b_t = (1 + r)b_{t-1} + d(1 + r) \tag{1}$$

if the real rate of interest is constant at the value r. Thus the bond stock will in this case grow without bound; if $d > 0$, $b_t \to \infty$ as $t \to \infty$.

Barro (1976) has argued that under these circumstances the Ricardian equivalence argument breaks down. In particular, he suggests that the rate of growth of the bond stock cannot exceed the economy's rate of output growth—here temporarily taken to be zero—because "the value of the outstanding stock of debt at any point in time is bounded by the government's . . . present value of future taxing capacity" (1976, p. 343, n. 2).[5] In a similar vein, Sargent and Wallace (1981) have argued that "if the interest rate on bonds is greater than the economy's growth rate, the real stock of bonds will grow faster than the size of the economy. This cannot go on forever, since the demand for bonds places an upper limit on the stock of bonds relative to the size of the economy" (1981, p. 2).[6]

These arguments do not, however, seem entirely convincing. First, if the bond-issuance policy is permanently maintained, then taxes will never have to be collected to retire the debt so the difficulty concerning taxing capacity may not be decisive. And under the Ricardian view, government bonds are not regarded as net wealth to the private sector, so the size of the bond stock seems to be potentially irrelevant.[7] In any event, it would appear that a more formal analysis might be worthwhile. As it happens, that conjecture is verified in the following section: while the basic idea of the Barro and Sargent-Wallace conten-

[5] Taxing capacity in each period is assumed by Barro to be some exogenously given fraction of current output.

[6] This assumption, it should be noted, plays an important role in the Sargent-Wallace analysis.

[7] Irrelevance of continued bond growth is assumed, without any utility-maximizing justification, in McCallum (1978).

tions receives support from the theory, the precise statements concerning bond demand limitations do not.

III. Analysis

Our first task is to specify a maximizing model that incorporates the crucial component of the Ricardian view, namely, infinite-lived agents who correctly take account of the government budget constraint.[8] The model must also be one that accommodates money, bonds, and a physical asset. To that end, let us adopt a discrete-time, perfect-foresight version of the well-known model of Sidrauski (1967), modified to include government bonds.[9] In order to keep matters as simple as possible, let us first consider a version with no depreciation or population growth.[10]

Formally, we imagine an economy composed of a large number of similar households, each of which seeks at period t to maximize

$$u(c_t, m_t) + \beta u(c_{t+1}, m_{t+1}) + \beta^2 u(c_{t+2}, m_{t+2}) + \ldots \qquad (2)$$

Here c_t is consumption in period t and $m_t = M_t/P_t$, with M_t the household's nominal money stock at the start of t and P_t the price of the (single) consumption good in t. The within-period utility function u is assumed to be "well-behaved"—see Sidrauski (1967, p. 535)—so unique, positive values are chosen for c_t and m_{t+1}, $t = 1, 2, \ldots$ The discount factor β equals $1/(1 + \delta)$, with the time-preference parameter δ positive.

Each household has access to a production function that neous of degree one in its inputs, labor and capital. But supplied inelastically, this function can be written as $f(k_t)$. the stock of capital held at the start of t. The function f is satisfy the conditions $f' > 0$, $f'' < 0$, $f'(0) = \infty$, and $f'(\infty)$ – unique, positive value will be chosen for k_t in each period.

---

[8] It should be pointed out that I am using the term "Ricardian" to refer primarily to these characteristics of the private economy. An entirely different use of the term occurs in Sargent (1982, p. 385), where it refers to a type of government policy regime—one that does not encompass policy variable paths such as those in our examples.

[9] The Sidrauski model is, of course, one in which bonds and money can easily coexist because real money balances appear as an argument of the household's within-period *utility function*. The rationale for this appearance—which has been severely criticized by Bryant and Wallace (1980) and others—is that transaction costs are reduced by money balances, so that more preferred bundles of consumption and leisure can be obtained. For a rather lengthy discussion of related issues, see McCallum (1983).

[10] A notable feature of the Sidrauski model is the invariance of the steady-state capital-labor ratio to expected *inflation* rates. This invariance (or superneutrality) does not survive minor modifications, such as making utility dependent upon leisure. It is my *impression* that the superneutrality property is not crucial for the issues under discussion here; the superneutral version of the model has been adopted for simplicity

exclusive of interest payments on current debt. More specifically, we can show that the model at hand will not support a zero-inflation equilibrium in which a permanently maintained positive deficit of $g_t + v_t = d$ is financed entirely by bond sales. To do so, we first observe that this monetarist hypothesis implies that $(1 + \pi_t)m_{t+1} - m_t = 0$ and that steady-state conditions prevail for all variables except b_t so that r_t and λ_t are constant, with $r = \delta > 0$ and $\lambda > 0$.[12] Then we insert the hypothesized constant values in (12), obtaining

$$b_{t+1} = (1 + r)b_t + (1 + r)d, \quad t = 1, 2, \ldots . \tag{14}$$

Next we note that the latter implies

$$b_{t+1} = (1 + r)^t b_1 + (1 + r)d[1 + (1 + r) + \ldots + (1 + r)^{t-1}], \tag{15}$$

which in turn implies

$$b_{t+1}\beta^{t-1}\frac{\lambda_t}{(1 + r_t)} = \lambda b_1 + \frac{\lambda d[(1 + r) - (1 + r)^{1-t}]}{r}. \tag{16}$$

But it can then readily be seen that the expression in (16) approaches $\lambda b_1 + [\lambda d(1 + r)/r]$ as $t \to \infty$, which violates condition (10). Equivalently, but in different words, b_{t+1} grows at the rate r while β^{t-1} decays at the rate $\delta = r$, so their product grows at the rate zero—that is, does not vanish as t increases. Thus one of the jointly sufficient conditions for household optimality is not satisfied by the proposed path. Now, since transversality conditions are not in all cases necessary for optimality,[13] we must also show that the proposed path is in fact inconsistent with optimality. To do so, we note from (12) that $(1 + r)^{-1}b_{t+1} - b_t = d$ is positive in each period, $t = 1, 2, \ldots$. Then inspection of (3) indicates that by setting $b_{t+1} = 0$ the household could increase its consumption in each period, $t = 1, 2, \ldots$, without changing its holdings of capital or real money balances. Thus the proposed monetarist path cannot be an equilibrium.

Our next object is to show that a constant, positive deficit can, however, be financed entirely by bond sales with no resulting inflation if "deficit" is defined—as it typically is—to include current interest payments. To make this argument, let us define the issue value of bonds outstanding at t by $\tilde{B}_t = B_t/(1 + R_{t-1})$. Then with $M_{t+1} - M_t = 0$, the government budget identity becomes

$$\tilde{B}_{t+1} - \tilde{B}_t = P_t(g_t + v_t) + \tilde{B}_t R_{t-1}. \tag{17}$$

---

[12] In this experiment (and those that follow) it is assumed that steady-state values for k and m prevail in period 1. Since the crucial aspects of the analysis involve limiting conditions, this simplification seems unobjectionable for the issues at hand.

[13] An example in which the transversality condition is not necessary for optimality was mentioned by Arrow and Kurz (1970, p. 46).

Now assume that policy keeps the real value of the right-hand side of (17) constant at $\bar{d}$, so that

$$\bar{B}_{t+1} - \bar{B}_t = \bar{d}P_t. \tag{18}$$

Next we conjecture that, with $\bar{d}$, g, and M held constant, the price level will also be constant so that (18) becomes

$$b_{t+1} - b_t = \bar{d}(1 + r). \tag{19}$$

But in this case we have

$$b_{t+1} = b_1 + \bar{d}(1 + 1 + \dots + 1^{t-1})(1 + r) = b_1 + \bar{d}t(1 + r), \tag{20}$$

so that

$$b_{t+1}\beta^{t-1}\frac{\lambda_t}{(1 + r_t)} = \frac{\lambda b_1 + \lambda\bar{d}t(1 + r)}{(1 + r)^t}, \tag{21}$$

which does approach zero as $t \to \infty$.

To verify that a noninflationary steady state also satisfies the other relevant conditions, we argue (constructively) as follows. First, with constant m and $\pi = 0$, $v_t = (1 + r)^{-1}b_{t+1} - b_t - g$ by (12), so (3) reduces to (13). Thus, given values for $\bar{d}$ and g, steady-state values of k, m, c, and λ are determined by (4), (5), (6), and (13). With $b_t > 0$ for all t, then, equation (7a) holds as an equality and determines $r > 0$. Finally, conditions (8) and (9) are satisfied with constant values for m, λ, and k since $\beta^{t-1} \to 0$ as $t \to \infty$.

The policy rule in the last example is one that makes b grow
rate that decreases over time, approaching zero in the
next consider an example in which b_t grows at a constant
that is numerically smaller than δ. For simplicity, let us
the rate is $\delta/2$. From our previous discussions it is clear that
condition (10) is not violated. Furthermore, it can be
that conditions (3)–(9) are all satisfied by constant values
λ, and $\pi = 0$. The behavior of v_t in this case satisfies

$$\begin{aligned} g + v_t &= (1 + r)^{-1}\left(1 + \frac{\delta}{2}\right)b_t - b_t \\ &= \left[\left(1 + \frac{\delta}{2}\right)(1 + \delta)^{-1} - 1\right]b_t = -\left(\frac{\delta}{2}\right)(1 + \delta)^{-1}b_t, \end{aligned} \tag{22}$$

so $d_t = g + v_t$ is negative and decays at the rate $\delta/2$. The alternative (conventional) deficit measure $\bar{d}_t = d_t + \delta b_t$ is positive, however, and grows at the rate $\delta/2$. Thus we see that the conventionally defined deficit can—in the Ricardian/monetarist Sidrauski model—grow forever without causing inflation. Furthermore, we see that the real stock of bonds can grow forever at a rate exceeding zero—which is, in this instance, the growth rate of the economy.

From these three examples, then, we obtain the following conclusions regarding bond finance of deficits in the basic, zero-growth version of our Ricardian/monetarist economy: (i) a permanent deficit cannot be financed solely with bonds if the deficit is defined exclusive of interest payments; (ii) a permanent deficit can be financed solely with bonds, and without inflation, if the deficit is defined inclusive of interest payments; and (iii) a positive growth rate of bonds outstanding can be maintained permanently, but only if the growth rate is smaller than the rate of time preference.

IV. Extensions and Qualifications

The purpose of this section is to describe both extensions and qualifications to the three basic conclusions above. We begin by developing extensions involving population growth and depreciation and then turn to some considerations not reflected in the formal model—default incentives, for example—that have the effect of limiting the applicability of the basic conclusions.

First, let us suppose that—as in Sidrauski (1967)—the size of each household grows at the rate n and that the utility function remains as in (2) with c_t and m_t now measuring per capita values. Then, with all other quantities also expressed in per capita terms, the household budget constraint becomes

$$\begin{aligned} f(k_t) + v_t = c_t &+ (1 + n)(1 + \pi_t)m_{t+1} - m_t \\ &+ (1 + n)(1 + r_t)^{-1}b_{t+1} - b_t + (1 + n)k_{t+1} - k_t. \end{aligned} \tag{3'}$$

One relevant effect is that the counterpart of (7) now implies that $(1 + n)(1 + r_t)^{-1} = (1 + \delta)^{-1}$ if $b_t > 0$ and $\lambda_t = \lambda_{t+1}$. Another is that condition (10) is replaced with

$$\lim_{t\to\infty} b_{t+1}\beta^{t-1}\frac{\lambda_t(1 + n)}{(1 + r_t)} = 0 \tag{10'}$$

In addition, the government budget identity becomes, in per capita terms,

$$(1 + n)(1 + \pi_t)m_{t+1} - m_t + (1 + n)(1 + r_t)^{-1}b_{t+1} - b_t = g_t + v_t. \tag{12'}$$

Consequently, if the per capita magnitudes g_t, v_t, and M_t are held constant over time—so that the aggregate deficit and money stock each grow at the rate n in every period—the following equation will govern the behavior of b_t under the monetarist hypothesis:[14]

$$b_{t+1} = (1 + r)(1 + n)^{-1}b_t + (d - nm)(1 + r)(1 + n)^{-1}. \quad (14')$$

Thus the per capita bond stock grows at the rate $(1 + r)(1 + n)^{-1} - 1$. But since $1 + r$ equals $(1 + n)(1 + \delta)$ in the hypothesized steady state, b_{t+1} then grows at the rate δ, which equals the rate at which β^{t-1} contracts. So the product $b_{t+1}\beta^{t-1}$ grows at the rate zero. This violates (10′) as in our first example with $n = 0$.[15] But it *just* violates (10′), so it is clear that the second and third examples, in which the per capita bond stock grows at a diminishing rate that approaches zero and at a constant rate less than δ, respectively, will not violate (10′). Thus the conclusions above hold in per capita terms when the economy experiences population growth of the specified type.

Furthermore, allowing proportional depreciation of capital would leave the crucial relationship between δ and the growth rate of bonds unchanged. The marginal product of capital would exceed r_t by the amount of the depreciation rate, but the steady-state condition $1 + r = (1 + \delta)(1 + n)$ would continue to hold and it is the relationship between r and **δ** that governs the relative growth rates of b_{t+1} and β^{t-1}.[16]

Now we turn to the qualifications. Given the rather surprising nature of conclusions ii and iii, it is important to recognize that some potentially significant aspects of actual economies are not reflected in the formal model. One of these involves the notion of "taxing ity," mentioned above. The suggestion put forth by B namely, that government debt cannot permanently gro excess of output growth, is based on the presumption th tions cannot forever grow faster than output. The te analysis that enables the third example to contradict th reasonable presumption is the recognition that househ able income includes interest payments from the govern as income from production. Thus taxes can exceed output for household and yet be smaller in magnitude than the household's disposable income. A reformulated version of Barro's suggestion, one stating that government debt cannot forever grow at a rate in excess of disposable income, would not be contradicted by any of our examples.

---

[14] Under that hypothesis, the inflation rate will be zero

[15] Again we avoid reliance on this transversality condition, for which necessity has not been established Instead we can show that the proposed path is nonoptimal by inferring from (3′) and (14′) that a zero value for b_{t+1} in each $t = 1, 2, \ldots$ would permit increases in c_t with the same paths for m_{t+1} and k_{t+1}

[16] Incomplete investigations with technical progress lead to the conjecture that the results remain valid in cases in which steady-state growth is possible

A Positive Model of Private Charity and Public Transfers

Russell D. Roberts

University of Rochester

This paper explores a model where private charity and public transfers are determined simultaneously. In political equilibrium, the government "overprovides" public transfers, transferring more to the poor than altruistic taxpayers prefer. At this equilibrium, private charity is zero. Evidence for this result is found by examining various types of data from the 1920s to the present. While private charity currently exceeds $50 billion, very little of it goes to the poor. I provide evidence that this phenomenon of zero private charity began, as the model predicts, in the 1930s, the beginning of federal intervention in the charity market.

I. Introduction

The growth in the size of government transfers is well known. At the same time, private charity reached an all-time high of $53.6 billion in 1981 (*New York Times*, April 9, 1982). While many authors have suggested that there is a negative relationship between public transfers and private charity, evidence for this crowding out has been sparse.

This paper derives a rigorous statement of the crowding-out relationship. Private charity is motivated by altruism. The political process then adjusts the distribution of income in order to maximize the political support received from both rich and poor. This simple model yields a number of strong predictions:

I am grateful to the following people for helpful comments and discussions: Gary Becker, Stanley Engerman, Matthew Goldberg, Bruce Jacobs, Walter Oi, Sherwin Rosen, Alan Stockman, and Michael Wolkoff. The usual disclaimer applies.

[*Journal of Political Economy*, 1984, vol. 92, no. 1]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0009$01.50

1. Private charity to the poor is reduced dollar for dollar by public transfers.
2. In political equilibrium, there is overprovision: there is more redistribution than is preferred by altruists.
3. In political equilibrium, public transfers reduce private charity to zero.

As evidence I show that the growth in public transfers to the poor during the Great Depression reduced private charity and caused a fundamental transformation of private agencies away from the relief of poverty toward other activities. This remains true today, with only a fraction of that $53.6 billion going to the poor.

Section II of this paper models private charity and its response to government policy. The inefficiency of private charity with many altruists is the impetus for government intervention. Efficiency can be achieved either by a system of tax credits or by direct taxation to finance public transfers. I focus on public transfers and their effect on private charity. Section III is an equilibrium model of the political process incorporating altruism. Section IV presents evidence on private and public transfers that is consistent with the theory. A final section concludes the paper.

II. A Model of Private Charity

Define altruism as the case where the level of consumption of one individual enters the utility function of the other. For example let A be altruistic and let C's consumption enter A's utility fun[illegible] sume for simplicity that C is not altruistic, so utility of A [illegible] written as

$$U_A = U_A(x_A, x_C)$$

$$U_C = U_C(x_C),$$

where x is the only consumption good. A's endowment is x_A an [illegible] assumed for simplicity to have an endowment of zero. If the act of transfer is costless, then A will set the marginal rate of substitution (MRS) between own consumption and that of C equal to one, the slope of the budget line. This is shown in figure 1 at E^*.

Now consider a world of two altruists, A and B, where each altruist takes the transfer of the other as given, the Cournot-Nash solution.[1] The equilibrium transfers T_A^* and T_B^* are such that, given T_B^*, T_A^* maximizes A's utility and T_B^* maximizes B's utility given that A's transfer is T_A^*. At this noncooperative equilibrium, E_1, the MRS of each

---

[1] See Goldberg (1979) for an analysis of the Stackelberg case

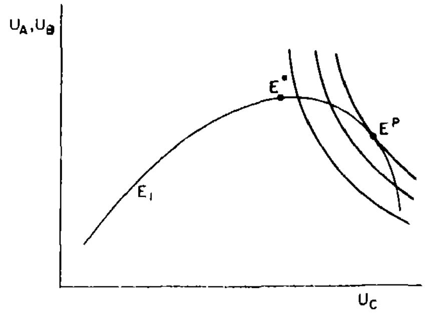


FIG. 3

redistribution is determined politically with the level of public transfers determined by the maximization of a function that has the utility levels of political constituents as arguments:

$$P = P(U_A, U_B, U_C). \tag{5}$$

Thus the political power of individuals is crucial, but efficiency considerations are important because individuals' utility, which enters the function, depends on the consumption of others.

The first partials of (5) are assumed to be positive. This type of function is found in Becker (1978), and a related idea is in Peltzman (1976). The level of the function is an index of political support; competition among political agents ensures that it gets maximized. Maximizing (5) requires that the distribution of income be efficient.[8] Which efficient point of the infinite number available maximizes (5)?

Figure 3 shows the utility possibility frontier and a set of iso-support curves—the set of utility realizations that keeps political support constant. Political equilibrium is found at the tangency between the utility frontier and an iso-support curve. With A and B assumed to be identical, one picture is sufficient. Altruism gives the frontier a hill shape. At low levels of U_C increases in x_C make A, B, and C happier.

The noncooperative point, E_I, is found on the upward-sloping portion of the frontier. Increases in x_C past E_I continue to increase U_A, U_B, and U_C, until E^*, the altruists' preferred point, is reached. The efficient points include E^* and those to the right. Increases in x_C beyond E^* benefit C at the expense of A and B.

If C has no political power, then the iso-support curves are horizontal and E^P and E^* are identical. But if C has any political power, that is, if (5) has positive first partials, then the iso-support curves are downward sloping and E^P must lie to the right of E^*. At E^*, the increase in political support from C when C receives an extra dollar outweighs the loss in support from A and B. The intuition is that at E^* small increases in x_C leave U_A and U_B unchanged since they are at a maximum but C is better off. Eventually this gain is offset by the losses to altruists and E^P is reached. This marginal political power of the poor at E^* results in E^P being to the right of E^*.[9]

---

[8] This is not the same as saying that government passes only efficient laws. In a world of more complex policy options and costly redistribution, maximizing (5) could mean passing laws that had dead-weight loss associated with them.

IV. Implications for the Interaction between Public Transfers and Private Charity

This result has two strong implications. First, in equilibrium, at E^P, private transfers are zero. Private transfers are zero to the right of E_1. Since E^P must lie to the right of E^*, which is to the right of E_1, private transfers are zero at E^P. In the United States, spending on supplemental security income, food stamps, AFDC, and other government programs dwarfs private charity. Private charity in the United States is approximately zero. Measured charity by individuals was $11.5 billion in 1981. But most of these contributions are charity as defined by the Internal Revenue Service, not charity in the sense of helping the poor. They are mostly contributions to religion, education, health, and the arts.

Table 1 summarizes the findings of the 1982 study, *Giving, U.S.A.*, conducted by the American Association of Fund-raising Counsel. Private charity to the poor is in the social services category. Of the $53.6 billion donated to charity in 1981 by individuals, corporations, and foundations, only $5.32 billion was categorized as social serv only a fraction of that amount went to the poor. The soc category includes agencies such as the Salvation Army ar organizations such as the United Jewish Appeal, Catholic and the Federation of Protestant Welfare Agencies. The re gory includes only sacramental spending that does not go to

But the social services category also includes the Boy Scouts and donations to the YMCA and YWCA. Almost on the spending for social services in 1981 went to the Uni United Way spending varies by community but typically ma spending on health and other items that only partially help the poo such as day-care centers. So the social services figure, which was at

---

[9] Here I have used the concept of a support function to derive this result. More generally, it will hold for other models of the political process where preferences of different interest groups are given a nonzero weight in determining outcomes. The support function is just one way of capturing this property.

TABLE 1

Categories of Recipients of Individual, Corporate, and Foundation Donations, 1955–81 (in Billions of 1981 $)

Year	Religion	Education	Social Services	Health & Hospitals	Arts & Humanities	Civic & Public	Others
1955	11.33	2.49	5.20	2.03	.68	.26	.67
1956	12.54	2.13	4.75	3.08	.48	.23	.48
1957	13.01	2.81	4.85	3.27	.52	.26	.82
1958	13.09	3.08	4.64	3.60	.51	.26	.50
1959	14.25	4.19	4.47	3.59	.62	.28	.57
1960	14.69	4.60	4.32	3.47	.58	.29	.86
1961	15.23	4.77	4.47	3.59	.59	.30	.90
1962	15.77	5.04	4.72	4.08	.63	.31	.90
1963	16.56	5.73	5.07	4.39	.67	.34	1.01
1964	17.33	6.01	4.96	4.25	.60	.46	1.75
1965	18.76	6.51	5.36	4.21	.65	.50	2.30
1966	19.78	6.58	5.35	5.80	1.05	.64	1.96
1967	20.16	6.68	5.15	7.68	1.03	.63	1.66
1968	21.46	7.65	5.51	7.88	1.06	.60	1.69
1969	21.96	7.87	6.30	8.02	1.36	.97	2.09
1970	21.78	7.59	6.74	7.96	1.46	.97	2.09
1971	22.43	8.00	6.56	8.20	2.06	1.34	2.74
1972	21.78	7.96	6.50	8.20	2.02	1.41	2.78
1973	21.51	8.38	6.26	8.49	2.56	1.25	3.91
1974	22.03	7.64	5.76	7.92	2.40	1.42	3.97
1975	21.81	6.67	5.30	7.47	2.90	1.50	4.48
1976	22.72	7.47	5.05	7.75	3.69	1.71	3.90
1977	25.42	7.13	5.28	7.36	3.41	1.76	3.72
1978	25.73	7.76	5.53	7.64	3.49	1.65	3.54
1979	25.24	7.51	5.45	7.46	3.38	1.50	3.68
1980	24.45	7.37	5.22	7.16	3.27	1.50	3.72
1981	24.85	7.49	5.32	7.36	3.35	1.54	3.71

Source.—American Association of Fund-raising Counsel 1982.

virtually the same level in real terms in 1981 as it was in 1955, is an upper bound for measured transfers to the poor. This figure in 1981 was less than 10 percent of total private giving and an even smaller fraction of government spending on the poor, the exact fraction depending on how government programs are classified. Private transfers to the poor are simply not very large.

But the model says more than that private charity is zero. It predicts that private charity first became negligible when government first intervened in a significant way in the charity market. Significant government intervention began in the 1930s and has continued to grow over time.[10]

From the data to follow, a stylized picture of the 1930s emerges:

---

[10] There were public transfers before the 1930s, but the amounts were at the state and local levels, not the federal one.

TABLE 2

EXPENDITURES FOR RELIEF FROM PUBLIC AND PRIVATE FUNDS IN 120 URBAN AREAS, 1929–35 (in Thousands of 1929 $)

Year	Public Funds	Private Funds
1929	33,449	10,296
1930	56,158	10,944
1931	138,874	55,663
1932	315,061	71,619
1933	557,658*	36,939
1934	835,425*	18,804
1935	1,035,206†	14,536

SOURCE.—Geddes 1937.
* Excludes expenditures under the Civil Works Administration.
† Excludes expenditures under the Works Program.

private donations fell dramatically as public transfers rose. But they did not fall to zero. Instead, charitable donations underwent a fundamental transformation during the period. They became less concerned with poverty and more concerned with health and social counseling.[11]

Table 2 shows public and private transfers over time from 120 urban areas. From 1930 to 1932 both public and private transfers increased approximately sixfold. But while public transfers grew about 3½ times between 1932 and 1935, private transfers fell by a factor of 5 and approached the level of 1929. Unfortunately, the source for table 2 (Geddes 1937) stops in 1935. But other evidence suggests not only that the trend continued but that the uses of [illegible] charity changed dramatically over this period. The New [illegible] ation for Improving the Condition of the Poor (AICP) [illegible] in 1843. For 96 years it transferred resources to the po[illegible] ployed, the sick, the elderly, and the husbandless moth[illegible] 1843 and 1939 it was one of the two most important pr[illegible] in the city, the other being the Charity Organization So[illegible] brella for a number of smaller charities. Between the [illegible] 1930s, the AICP's role and resources changed dramatically, as [illegible] trated in table 3.

With the coming of the Depression, donations, total expenditures, and expenditures on material relief increased up until 1932.[12] This is the same pattern found in the national data. From 1932 to 1935, the experience of the AICP continues to parallel the national data—there

---

[11] Thus there is a serious problem with using aggregate donations to charity from 1929 onward. A dollar contributed to charity in 1929 bought a very different bundle of services than it did in 1939, e.g.

[12] Material relief is direct transfers of money and resources. Total expenditures include camps and homes run by the association for the poor, children, and the elderly.

TABLE 3

MATERIAL RELIEF, EXPENDITURES, AND DONATIONS, NEW YORK ASSOCIATION FOR IMPROVING THE CONDITION OF THE POOR, 1928–38 (in 1929 $)

Year	Material Relief	Total Expenditures	*Donations**
1928	570,347	1,320,885	900,680
1929	538,167	1,397,047	827,286
1930	910,946	1,848,467	883,012
1931	1,377,964	2,518,749	1,092,823
1932	1,675,220	3,100,696	1,999,996
1933	1,115,593	2,718,344	1,712,399
1934	1,177,580	2,003,945	1,589,210
1935	1,096,386	1,931,303	1,306,765
1936	690,450	1,913,188	789,072
1937	541,153	1,848,110	855,906
1938	556,118	1,895,265	602,679

SOURCE.—Annual Reports of the Association, 1928–38.
* Excludes legacies. It is difficult to track legacies during the period.

is a decrease in donations and expenditures. But the most dramatic changes are from 1936 to 1938 when national data are missing. Donations fall to one-third of their 1932 level and two-thirds of their level in 1928. These results are consistent with the model.

An alternative explanation of these data is that as donors' incomes fell during the Depression, charity decreased and government had to take up the slack. Evidence against this explanation is found in the depression of 1890–94. In 1890–94 public transfers in the city of New York grew only 15 percent. At the same time, expenditure by the AICP grew by a factor of 4 over the 4-year period (Newcomer 1941, p. 654). So, despite a decrease in income, private charity rose substantially when increases in public transfers were relatively small. Even between 1929 and 1932 (see table 2), private expenditures grew sevenfold nationally. Donations to the AICP continued to grow until 1932 (table 3). Only when public spending continued to grow did private spending go to virtually zero.

Of equal importance with the large drop in donations from 1932 to 1938 is the dramatic change in the activities of the AICP away from antipoverty work. While donations did not go completely to zero, evidence suggests that transfers to the poor did. In 1939 the association merged with the Charity Organization Society forming the Community Service Society. This merger made the trend away from fighting poverty official. The theme is a common one in the annual reports of the AICP in the 1930s: "The AICP has made major revisions of budgets downward. Many families formerly cared for by the AICP have been turned over completely to public relief departments. Nearly one-third of the present number of families under care are cooperative cases with public authorities, in which the cooperation consists in the AICP supplying social services not yet available in public departments" (Annual Report of the New York Association for Improving the Condition of the Poor, 1935–36, p. 10).

In 1939, *Survey Midmonthly,* a social work journal, reported the merger and explicitly recognized crowding out of private charity: "Both reports [by the two merging agencies] emphasize the freedom that has come to the private agencies with the public assumption of responsibility for relief. The AICP saw in this new found freedom a chance for the expansion of its health work, thus digging at the roots of one of the main causes of poverty and distress. The COS saw an opportunity for more intensive work in rehabilitation and family counsel" (*Survey Midmonthly,* May 1939, p. 152).

This shift from material relief to social counseling became more pronounced over time. Additional evidence suggests that this trend was a national one and that eventually private agencies came to ignore the poor, with counseling resources going to higher-income families. Dickinson (1970, p. 69) argues that religious organizations no longer dealt with the poorest families, letting these cases be handled by public agencies. Cloward and Epstein (1965) have documented this phenomenon across a wider set of organizations. They write, "Once publicly supported income maintenance programs came into existence, following the depression, private agencies began to refer economically deprived clients [to public agencies], thus conserving their resources for other services" (Cloward and Epstein 1965, pp. 623–24).

Initially, in the mid-thirties, public and private relief ag
intended to supplement each other. However, "this
labor—that is, private agency supplementation of public
. . . short-lived. As the pall of the depression lifted . . . a
tion of private casework began to emerge—one heavily d
psychological conceptions of family problems. It tended t
importance of environmental approaches (housing, e
medical, and other concrete environmental services) . . . The p
agency began to limit its responsibility for poor people to conducting studies and to giving expert testimony about the current needs of welfare recipients" (Cloward and Epstein 1965, pp. 624–25)

This evidence suggests an explanation for some of the puzzling findings in the research on the trade-off between public transfers and private charity. Using data from 1948–72, Abrams and Schmitz (1978) found that a dollar increase in government welfare expenditures reduced private charity by 28 cents. Schwartz (1966) reports that his estimates over 1929–60 resulted in a small but *positive* relationship between public and private charity. In both of these studies itemized charitable deductions from income tax returns are used as the dependent variable. Since much of these data consist of religious contributions, it is not surprising that a dollar-for-dollar trade-off or even a negative one is not found.[13]

The evidence presented raises questions that the model is too general to answer. Public transfers in the 1920s, though small by today's standards, were not zero, nor were private transfers. We would like to know how the composition of private agencies was altered before the 1930s as public funds at the state and local level began to be used for different types of relief—aid to husbandless mothers, the blind, the aged, and the unemployed. I suspect that private agencies were crowded out of each of these in turn as they were eventually crowded out of all of them. Obtaining data to test this idea would be extremely difficult, however.

The model of altruism presented here, while simple, is not a very rich picture of compassion. Altruism goes far beyond a desire to raise the consumption level of the least fortunate. As government takes over this task our compassion turns to imperfect substitutes: counseling services, alleviating the diseases that afflict children, and helping the less fortunate of other nations.

Another empirical application of the model is to the recent cuts in social welfare spending proposed by the Reagan administration. The model predicts that private charity will not increase when government cuts welfare spending. Changes in public transfers are movements from one equilibrium point like E^P to another. Private transfers are zero at either point. This does not imply a lack of compassion on the part of Americans. The only time a trade-off between private charity and public transfers is observed is when government moves from E_1 to E^P, as occurred at the federal level in the 1930s or today if for some reason public transfers were reduced to zero. The model also predicts that the large cuts originally proposed by Reagan for 1984 and beyond will not be enacted since it is unlikely that there have been large changes in the tastes of altruists or political power, the determinants of public transfers in figure 3. The actual cuts have been relatively small.

Another implication of political equilibrium is overprovision. The attempt of the altruists to avoid E_1 and achieve E^* by putting redistribution into the political arena is partially frustrated by the marginal political power of the poor at E^*. This is a cost associated with living in a society with representative government. This is a price worth paying unless C's political power is unreasonably large, since altruists are better off than they were at E_1. However, individual altruists could easily be worse off if they were less altruistic than others or if their political power were less than that of other altruists and it was sufficiently inexpensive to use policies that discriminated against them.

---

[13] Schwartz says, "A predominant part of private donations, however, go to religious causes, while no government funds of the welfare category are so channeled. . . . Thus, similarity of titles gives a spurious feeling of similarity of purpose, the two series are, to a marked extent, not comparable" (1966, p. 34).

On the other hand, altruists have incentives to disguise their preferences. An individual who is indifferent toward the poor would pretend to dislike welfare recipients to reduce his tax burden. Though I assume full information on the part of the political process with respect to preferences, this phenomenon may be important. It suggests that professed views and survey responses on the level of government transfers understate the altruism felt by taxpayers.

V. Conclusion

This paper integrates two explanations for the level of public transfers—the public good argument and the political power of recipients. One result that emerges is that the political process "overprovides" public transfers in the sense that more resources are transferred to the poor than altruists desire. This implies that private charity is zero.[14]

Current data and evidence from the Depression yield support for the crowding-out result.[14] The huge growth in public transfers in the 1930s crowded out private antipoverty efforts and fundamentally changed the nature of private charity. Current data also support this conclusion.[15]

As usual, this paper uses a number of simplifying assump
Public charity is a perfect substitute for private charity su
and recipients only care about the consumption level of
altruists receive utility from the act of transferring as we
vate charity would still be positive in equilibrium. The
concentrates on homogeneous altruists. A more intensi
tion of a world of heterogeneous altruists may lead to furt
tions and help explain why some goods are publicly provide
others are not.

---

[14] Since private charity is also zero at E^*, zero private charity does not imply overprovision.

[15] The evidence for the model came from the United States. England has a much older tradition of public provision for the poor. Of England's experience, Poynter (1969, p. 3) has written, "it is apparent that in the century after 1660 payments under the Poor Law became almost everywhere the ordinary source of relief for indigence, with private charity a supplementary source of varying importance, called on for great efforts only in times of extraordinary distress." I thank Stan Engerman for bringing this source to my attention.

References

Abrams, Burton A., and Schmitz, Mark D. "The 'Crowding-Out' Effect of Governmental Transfers on Private Charitable Contributions." *Public Choice* 33, no. 1 (1978): 28–40.

American Association of Fund-raising Counsel. *Giving, U.S.A.* 1982.

Becker, Gary S. "A Positive Theory of the Redistribution of Income." Working Paper no. 006. Chicago: Univ. Chicago, Center Study Econ. and State, April 1978.

———. "A Theory of Competition among Pressure Groups for Political Influence." *Q.J.E.* 98 (August 1983): 371–400.

Cloward, Richard, and Epstein, Irwin. "Private Social Welfare's Disengagement from the Poor: The Case of Family Adjustment Agencies." In *Social Welfare Institutions: A Sociological Reader*, edited by Mayer N. Zald. New York: Wiley, 1965.

Dickinson, Frank G. *The Changing Position of Philanthropy in the American Economy*. New York: Columbia Univ. Press (for Nat. Bur. Econ. Res.), 1970.

Geddes, Anne E. *Trends in Relief Expenditures, 1910–1935*. Research Monograph X. Works Progress Administration, Division of Social Research. Washington: Government Printing Office, 1937.

Goldberg, Matthew. "Managerial Discretion and Corporate Contributions to Charity." Manuscript. Chicago: Univ. Chicago, Dept. Econ., January 1979.

Hochman, Harold M., and Rodgers, James D. "Pareto Optimal Redistribution." *A.E.R.* 59 (September 1969): 542–47.

Newcomer, Mabel. "Fifty Years of Public Support of Welfare Functions in the United States." *Soc. Service Rev.* 15 (December 1941): 651–60.

New York Association for Improving the Condition of the Poor. Annual reports.

Peltzman, Sam. "Toward a More General Theory of Regulation." *J. Law and Econ.* 19 (August 1976): 211–40.

———. "The Growth of Government." *J. Law and Econ.* 23 (October 1980): 209–87.

Poynter, John R. *Society and Pauperism: English Ideas on Poor Relief, 1795–1834*. Toronto: Univ. Toronto Press, 1969.

Roberts, Russell. "Altruism, Efficiency and the Redistribution of Income." Manuscript. Rochester, N.Y.: Univ. Rochester, Dept. Econ. and Public Policy Analysis Program, July 1983.

Schwartz, Robert A. "Private Philanthropic Contributions—an Economic Analysis." Ph.D. dissertation, Columbia Univ., 1966.

Survey Midmonthly. May 1939.

Warr, Peter G. "Pareto Optimal Redistribution and Private Charity." *J. Public Econ.* 19 (October 1982): 131–38.

Comments

Misconceptions about the Real-Bills Doctrine: A Comment on Sargent and Wallace

David Laidler

University of Western Ontario

I

In their (1982) paper "The Real-Bills Doctrine versus the Quantity Theory: A Reconsideration," Sargent and Wallace develop certain properties of Samuelson's (1958) overlapping-generations-consumption-loan model under alternative monetary policy regimes. They argue that their analysis casts new light on the well-known debate to which they refer in their title, claiming that their results call for "something of a rehabilitation of the real-bills doctrine" (p. 1214) Their analysis per se has all the characteristics we have come to expect from these authors: it is elegant and rigorous, yielding concl
provocatively strong to the point of being paradoxical H
gent and Wallace's attempted rehabilitation of the real-
suffers from the following deficiencies: (1) the conclusio
basis they seek to rehabilitate the real-bills doctrine wou
anathema to its proponents; (2) what they refer to as
doctrine is not the real-bills doctrine; and (3) their inte
Adam Smith's analysis of the social productivity of bank
misconceived. I shall take up these points in turn.

II

Sargent and Wallace show that what they call a "free-banking or real-bills regime" (p. 1214) will lead to a fluctuating and perhaps indeter-

---

I am grateful to Robert Barro, Peter Howitt, Michael Parkin, George Stigler, Lawrence H. White, and two anonymous referees for helpful comments on earlier versions of this essay.

[*Journal of Political Economy*, 1984, vol. 92, no. 1]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0003$01.50

minate price level or perhaps to no monetary equilibrium at all, but that such a regime will also lead to a Pareto-optimal allocation of resources. The latter result, to them, justifies their defense of the regime, because their model provides no grounds for attaching any welfare significance to price-level instability or indeterminacy per se (see pp. 1233–34). More recently, Wallace (1982) has cited Sargent and Wallace (1982) in support of his contention that "so far as we know, the claim that a stable price level is good is as naive as the claim that a stable relative price between, say eggs and steel is good. In fact, it has been argued that the two are on the same footing" (p. 229). Such a conclusion is indeed implicit in Sargent and Wallace's model, and they tell their readers, correctly, that the main criticism leveled at the real-bills doctrine by its opponents has been that, when implemented, the doctrine leads to price-level instability. They fail to note, however, that eighteenth- and nineteenth-century advocates of the doctrine regarded the maintenance of price-level stability as a vital principle of social organization and would not have entertained a defense of their doctrine that showed that it failed to ensure such stability.

Adam Smith, to whom Sargent and Wallace pay particular attention, was one of the originators of the real-bills doctrine, and his views are typical in this respect.[1] Smith, being an eighteenth-century economist, did not talk in terms of such standard devices of twentieth-century monetary theory as "money in the utility function," "money in the production function," or "transactions technologies," which Sargent and Wallace dismiss as arbitrary and therefore irrelevant to their concerns (see pp. 1233–34). However, he certainly regarded the monetary system as being of the utmost importance for economic well-being, likening the stock of money to a piece of fixed capital (a water wheel and a waggon-way were among his analogies) that enhanced the productivity of labor by extending the market, thereby promoting the division of labor. Though he did not comment explicitly on the desirability of a situation in which there was no monetary equilibrium, it is safe to say that, given his views about the productivity of the monetary system, he would not have approved of an institutional framework that might lead to such a state of affairs.

Smith did discuss conditions of fluctuating, and in particular rising, prices with strong disapproval. Unlike Sargent and Wallace, he did not in his analysis attribute perfect foresight to agents, and he took careful account of the fact that, in the world around him, there existed long-term contracts denominated in money. Smith therefore understood the potential of price-level fluctuations for redistributing wealth in an arbitrary fashion, and he feared that potential: words such as "injustice," "fraudulent," and "cheat" enter his vocabulary when he discusses these matters (see, e.g., p. 350). Such sentiments are not confined to Smith. All subsequent important adherents of the real-bills doctrine held similar views. Thus, when Thomas Tooke, a leading figure in the banking school and a real-bills advocate, argued that the maintenance of the convertibility of currency into specie at a fixed price was "the *sine qua non* of any sound system of currency" (1840, p. 177), he was expressing a view that had been a commonplace on both sides of the real-bills–quantity theory debate for four decades and was to remain so well into the twentieth century.[2]

---

[1] I have dealt with Adam Smith's monetary economics in a recent paper (Laidler 1981). My treatment of his analysis in this Comment is of necessity brief, and the reader who is looking for a fuller account of it, well substantiated with quotations from the *Wealth of Nations*, is referred to that paper. The main body of Smith's monetary economics is to be found in bk. 2, chap. 2 of the *Wealth of Nations*, but relevant material is also to be found in bk. 1, chap. 5.

Sargent and Wallace assert that the real-bills doctrine "relies on market forces to prevent excessive 'credit creation' by private banks" (p. 1213). However, they do not point out that advocates and opponents alike of the real-bills doctrine agreed about what constituted "excessive credit creation." For them, when paper was convertible into specie, credit creation that led to a drain of specie from the reserves of the banking system, and hence endangered the maintenance of convertibility, was excessive. When paper was not convertible, credit creation by the banking system that led to a rising price level was excessive. The real-bills doctrine is also known as the real-bills *fallacy* precisely because its advocates claimed that the ope
procedures they advised would be adequate for avoid
credit creation thus defined when in fact they were not
Wallace's results, showing that their laissez-faire bankin
lead to a fluctuating and perhaps indeterminate price lev
no monetary equilibrium at all, far from providing the
rehabilitation of the real-bills doctrine, therefore in fact
fallaciousness; or rather, they would do so if Sargent and Wa
version of laissez-faire banking was the same thing as a real-bills regime; but it is not, as I shall now argue.

---

[2] Of course there were always, as there are nowadays, inflationary cranks who were unconcerned about price stability. In Tooke's day the Birmingham school sometimes took this position. The reader's attention is drawn to the fact that, until the discovery of the Californian and Australian gold fields in the middle of the nineteenth century, economists discussing the determinants of the general price level would often couch their discussion in terms of the paper money price of specie, because stability of the price of goods in terms of specie was usually taken for granted. During the Restriction Period (1797–1821), however, one element in monetary debates concerned the extent to which the high price of gold bullion reflected a rise in the general price level, and how far it reflected an increase in the price of gold relative to goods in general brought about by wartime conditions.

III

Sargent and Wallace tell their readers that the real-bills doctrine "proposes that there be unrestricted discounting of . . . evidences of indebtedness, which . . . are . . . free of default risk . . . . The key prescription of the doctrine is that no government regulations ought to restrict the scope of such intermediation" (pp. 1212–13).

It is instructive to compare this assertion with Smith's statement of the doctrine, part of which Sargent and Wallace quote (p. 1213, n. 1), but which I here reproduce in full.

> When a bank discounts to a merchant a real bill of exchange drawn by a real creditor upon a real debtor, and which, as soon as it becomes due, is really paid by that debtor; it only advances to him a part of the value which he would otherwise be obliged to keep by him unemployed and in ready money for answering occasional demands. The payment of the bill, when it becomes due, replaces to the bank the value of what it had advanced, together with the interest. The coffers of the bank, so far as its dealings are confined to such customers, resemble a water pond, from which, though a stream is continually running out, yet another is continually running in, fully equal to that which runs out; so that, without any further care or attention, the pond keeps always equally, or very nearly equally full. Little or no expense can ever be necessary for replenishing the coffers of such a bank. [Smith 1776/1966, p. 325]

Note that Smith does not advocate "unrestricted discounting of . . . evidences of indebtedness, which . . . are . . . free of default risk." Rather, he proposes that by restricting its discounting to a narrow subset of such evidences of indebtedness, the bank will avoid a drain on its "coffers" (reserves of specie), a drain that would threaten the convertibility of its liabilities. Note also that Smith suggests that bank lending be confined to a certain class of customers, namely, merchants, who should be thought of as borrowing to finance goods in the process of production and distribution. As an empirical matter, Smith did not believe that a banking system left unregulated would confine itself to so narrow a range of business. His insistence, which Sargent and Wallace note (p. 1213, n. 1), that banks be forced by law to issue only liabilities convertible on demand into specie was designed precisely to help guarantee the price-level stability that an unregulated banking system, as he believed, could not be relied on to provide.

Smith is not the only advocate of legal restrictions on banking among protagonists of the real-bills doctrine. Torrens, its principal exponent during the restriction period (1797–1821) in Britain, is recorded by Mints (1945) as having argued that, although under convertibility, restrictions on bank lending per se would be unnecessary, under an inconvertible regime such as ruled at the time, legislation was required to ensure that "the operations of banking are confined within their proper limits." The proper limits in question involved discounting only good short-term bills (see Mints [1945, p. 50], where Torrens's views are explained and the phrase above from his writings is quoted). Similar views, particularly about the need for a legal requirement of convertibility, are to be found in the contributions of the banking school to the debate about Sir Robert Peel's 1844 Bank Charter Act.[3]

In short, it was of the very essence of the real-bills doctrine that, far from bank lending being unrestricted, it should be confined to loans made on the security of short-term bills of exchange issued by reputable merchants or manufacturers to finance the production and distribution of real goods. The belief that bank lending thus confined would guarantee the preservation of price-level stability and/or specie convertibility was an integral, and fallacious, element of the doctrine. Furthermore, proposals to put legal limits on banks' activities, either by requiring specie convertibility on demand of their notes or, in its absence, by restricting the class of bills eligible for discount, were an integral part of the policy proposals associated with the real-bills doctrine. That doctrine did not, as Sargent and Wallace would have it, involve a defense of laissez-faire in the banking industry. In any event, a model in which banks are permitted to make unlimited *consumption* loans, the only kind that exist in the world analyzed by Sargent and Wallace, surely cannot be used to rehabilitate [illegible] whose central feature, as we have now seen, was that ban[illegible] restricted to short-term loans to finance current *produc*[illegible] *tion*, only one of the many kinds of loan available in the [illegible] lyzed by Smith and subsequent real-bills advocates.

IV

Sargent and Wallace analyze a laissez-faire banking system with no convertibility obligations and also one in which bank paper is convertible at a fixed price into commodity money, noting correctly that Smith's analysis of the productivity of banking is carried on in the context of the latter system. In discussing this matter, they attribute to Smith views he did not hold. Specifically, they say, "Smith argued that, through substitution of paper liabilities for gold in lenders' portfolios, the spread of banking would provide more favorable terms for borrowers. . . . We regard Smith as arguing that interest rates on loans would be lower and less gold would be held if banks were left largely unregulated" (pp. 1228–29). These conclusions follow from Sargent and Wallace's model but are only tenuously related to Smith's views, even as expressed in the passages from the *Wealth of Nations* that Sargent and Wallace quote. These passages make no reference to the rate of interest or to the terms available to borrowers.

---

[3] Detailed arguments to support the interpretation of Smith above are to be found in Laidler (1981). I have discussed the banking school–currency school debate in my essay "Thomas Tooke on Monetary Reform" reprinted as chap. 11 of Laidler (1975). Detailed and documented support for some of the positions I take in this Comment are to be found there.

Smith's theory of profits, and hence of interest, was not well developed in comparison with that of later English classical writers, but an acceptable summary of it would say that the supply and demand for real capital determined the rate of interest. Smith was emphatic that, despite superficial appearances to the contrary, interest was not a monetary phenomenon. "Almost all loans at interest are made in money, either of paper, or of gold and silver. But what the borrower really wants, and what the lender really supplies him with, is not the money, but the money's worth, or the goods which it can purchase. . . . the money is, as it were, but the deed of assignment, which conveys from one hand to another those capitals which the owners do not care to employ themselves" (pp. 376–77).

For Smith, banking enhanced the economy's wealth, not because it lowered the rate of interest by allowing agents to hold default-free claims to future goods rather than the stocks of goods themselves, as it does in Sargent and Wallace's model, but because, in replacing gold with paper in circulation, it left the monetary system as productive as ever—though not quite as safe—while permitting gold to be exported in exchange for wage goods. These in turn could be used to support a larger and more productive labor force.

The effects that Sargent and Wallace describe certainly follow from their own model, but in that model, population is exogenous and there is no production. Such a model cannot, as Sargent and Wallace claim, permit "a formal representation of arguments that Adam Smith advanced" (p. 1227). These hinged on the endogenous response of the size of the working population, and hence of output, to an increase in the stock of circulating capital, whose purchase on world markets was made possible by the replacement of gold by paper money in domestic circulation. In Smith's model such effects come about with no permanent fall in the rate of interest, because that model is not closed, as is, for example, Ricardo's (1821/1951) development of it, by diminishing returns in agriculture.[1] It is true that, in terms of the logic of his own model, Smith could have argued that an increase in the stock of wage goods brought about by an export of gold might *temporarily* raise wages and depress profits and hence interest, during the time it took the population's size to respond to this larger stock of circulating capital and in which therefore the supply of capital moved ahead of the demand. As far as I know, however, Smith did not anywhere suggest that the replacement of specie money with bank money would have even such a temporary effect on the interest rate, let alone the permanent effect that it has in Sargent and Wallace's model.

---

[1] The mechanism whereby, according to Smith, this export of gold took place was in some respects similar to the real balance effect that lies at the heart of the modern monetary approach to balance of payments analysis. On this, see Laidler (1981), pp. 188–92, see also Bloomfield (1975), p. 480.

V

If economics is to make progress, it clearly needs new analytic results, and Sargent and Wallace have certainly provided these. However, new results are only a necessary condition for progress, which also requires that, in the process of working out new ideas, we do not lose sight of or distort old ones. I hope that this note has done something to help ensure that, in the process of absorbing Sargent and Wallace's new and provocative analysis, readers of the *Journal of Political Economy* do not lose sight of what was at stake in earlier monetary debates.

References

Bloomfield, Arthur I. "Adam Smith and the Theory of International Trade." In *Essays on Adam Smith*, edited by Andrew S. Skinner and Thomas Wilson. Oxford: Clarendon, 1975.

Laidler, David. *Essays on Money and Inflation*. Chicago: Univ. Chicago Press, 1975.

——— "Adam Smith as a Monetary Economist." *Canadian J. Econ.* 14 (May 1981): 185–200.

Mints, Lloyd W. *A History of Banking Theory in Great Britain and the United States*. Chicago: Univ. Chicago Press, 1945.

Ricardo, David. *On the Principles of Political Economy and Taxation*. 3d ed. Reprinted in *The Works and Correspondence of David Ricardo*, edite Sraffa. Cambridge: Cambridge Univ. Press (for Royal Econ. Soc

Samuelson, Paul A. "An Exact Consumption-Loan Model of Inter Without the Social Contrivance of Money." *J.P.E.* 66 (Decem 467–82.

Sargent, Thomas J., and Wallace, Neil. "The Real-Bills Doctrine versus Quantity Theory: A Reconsideration." *J.P.E.* 90 (December 1982): 1212–36.

Smith, Adam. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 1776. Reprint ed., edited with an Introduction by Edwin Cannan. New Rochelle, N.Y.: Arlington House, 1966.

Tooke, Thomas. *A History of Prices, and of the State of the Circulation, in 1838 and 1839*. Vol. 3. London: Longman, Orme, Brown, Green, & Longman, 1840.

Wallace, Neil. "Panel Discussion." In *Interest Rate Deregulation and Monetary Policy*. San Francisco: Federal Reserve Bank of San Francisco, 1982.

Inferior Forecasters, Cycles, and the Efficient-Markets Hypothesis: A Comment

Hersh M. Shefrin

University of Santa Clara

I

In an interesting article, Figlewski (1978) discusses a model with intriguing implications for the long-run behavior of real-world speculative markets. The present note extends his analysis to study four specific aspects of speculative market dynamics.

1. Consider a trader whose market forecasts are generally inferior to those of the other traders. (One might think of somebody who finds himself trading against insiders.) With what likelihood will an inferior forecaster be able to achieve varying degrees of success in the market over time?
2. With what likelihood will an inferior forecaster actually be able to persist in the market indefinitely?
3. Does Figlewski's analysis suggest that there is any sense in which the strong form of the efficient-markets hypothesis is appropriate for speculative markets?
4. Do speculative markets have a tendency to move in cycles?

II

Consider a trader (A) who participates as a pure speculator in a sequence of spot and futures markets for a single commodity. Clearly A's success in choosing futures positions will depend on his ability as a forecaster. The first two issues taken up in this note concern the case in which A's forecasting ability is inferior to that of the other traders in the market. This might happen because the other traders have access to information not available to A (but *not* vice versa), or it might happen because A is less efficient at generating forecasts from common information. This does not preclude A from making more accu-

I would like to acknowledge the helpful remarks and suggestions of Stephen Figlewski, Sam Peltzman, and an anonymous referee

[*Journal of Political Economy*, 1984, vol. 92, no. 1]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0007$01.50

rate forecasts than the others in isolated instances, though such occurrences are to be regarded as rare. Nevertheless, it is of interest to have some idea about the likelihood with which an inferior forecaster will double or triple his wealth. To analyze A's prospects in this kind of setting, I shall employ Figlewski's (1978) two-trader framework. (For a treatment of the *N*-trader problem, see Figlewski [1982].)

Consider two traders, A and B, who participate in spot and futures markets for a single commodity over a sequence of periods. It is assumed that a futures market is held at the beginning of each period and a spot market is held at the end. Let the date t futures price be denoted by P_t and date t spot price by P_t^*. Suppose that trader A begins period t with wealth W_t^A and takes position n_t^A in the futures market. If $n_t^A > 0$, then A goes "long"; if $n_t^A < 0$, then A goes "short." We imagine that A's remaining wealth, $W_t^A - P_t n_t^A$, is invested in a risk-free security paying zero interest. Let the delivery date on a period t futures contract be the end of period t. Then trader A will use the spot market to earn a capital gain (or loss) of magnitude $n_t^A(P_t^* - P_t)$ on his position n_t^A.

In determining an appropriate position, A will have to take into account his wealth level W_t^A, the prevailing futures price P_t, his beliefs about the spot price P_t^*, which has yet to be determined, and his attitude to risk. Suppose that A's period t preferences are represented by a utility function that exhibits constant relative risk aversion (with risk parameter R^A) and whose arguments are the (subjective) mean and variance of end-of-period wealth. Moreover, suppose that A receives private information at the outset of t on which he bases a forecast P_t^A of the spot price P_t^*. Analogous remarks apply to trader B and the variables W_t^B, n_t^B, R^B, and P_t^B. For convenience, the sum of W_t^A and W_t^B will always be normalized to unity. Figlewski assumes forecast errors take the form

$$P_t^A - P_t^* = \epsilon_t^A + \delta_t; \quad P_t^B - P_t^* = \epsilon_t^B + \delta_t,$$

where $\epsilon_t^A \sim N(0, \sigma_A^2)$, $\epsilon_t^B \sim N(0, \sigma_B^2)$, $\delta_t \sim N(0, \theta^2)$, and ϵ_t^A are mutually independent. This assumption reflects the fact that A and B may have access to some of the same information (because δ_t is common to both) but allows for a degree of heterogeneity as well. Observe that for a given position n_t^A, A's subjective forecast of his end-of-period wealth W_{t+1}^A is given by

$$E_t^a(W_{t+1}^A | W_t^A) = W_t^A + n_t^A(P_t^A - P_t). \tag{2}$$

It follows from (2) that a utility-maximizing choice of n_t^A (and n_t^B) will depend on the futures price P_t. A futures market equilibrium will require that n_t^A and n_t^B sum to zero. Figlewski establishes that this will lead the equilibrium price P_t to be a convex combination of the forecasts P_t^A and P_t^B. Specifically,

$$P_t = X_t^A P_t^A + (1 - X_t^A) P_t^B, \tag{3}$$

where

$$X_t^A = \frac{W_t^A}{W_t^A + (R^B/R^A)(1 - W_t^A)}. \tag{4}$$

Significantly, the futures price P_t is an unbiased estimator of the spot price P_t^* (by [3]). I return to this point again in Section IV. Notice that the weighting factor X_t^A actually depends on W_t^A. One of Figlewski's principal assertions is that there is a unique constant X_{eq} such that X_t^A has the "tendency" to move toward X_{eq}. Therefore he calls the value of W_t^A that renders X_t^A equal to X_{eq} the "long-run equilibrium" value W_{eq}^A. Thus as these markets are repeated over time, a tendency exists for W_t^A to approach W_{eq}^A and W_t^B to approach $1 - W_{eq}^A$. Below I discuss the relationship between X_{eq} and the efficient-markets hypothesis.

III

Consider how well one might expect an inferior forecaster to do in Figlewski's market over time. In particular, consider the likelihood with which an inferior forecaster will depart from the market because his wealth dwindles down to zero. To address these questions, suppose that trader B has access to all of trader A's forecast information and more besides. Then the value added of A's forecast P_t^A to that of B will be zero in the sense that

$$E(P_t^*|P_t^A, P_t^B) = E(P_t^*|P_t^B) = P_t^B. \tag{5}$$

In this case A's forecast is Blackwell-inferior to that of B (see McGuire 1972).[1] Significantly, this does not rule out the possibility that in a given period A will earn a capital gain and B will incur a capital loss. Moreover, it does not even rule out the possibility that B runs up against a string of bad luck, loses all his wealth, and leaves the market. However, what I now argue is that this last occurrence is extremely rare so long as A's initial wealth W_0^A is small relative to total market wealth. Proposition 1 below strongly indicates that it is a virtual certainty that a small inferior forecaster will eventually be driven out of the market. The proposition also indicates that the probability of an inferior forecaster *ever* multiplying his initial wealth by a factor of K through speculation is no greater than $1/K$. Therefore proposition 1 applies to the first two issues described in Section I.

---

[1] If A's forecast variance is higher than that of B, but A has access to information not available to B, then we would not call A a Blackwell-inferior forecaster. In this kind of situation one expects that the special information in A's possession will enable him to persist in the market. Figlewski actually asserts that "traders with inferior information do not get driven out of the market by the better forecasters" (p. 591) because his definition of inferiority is based on forecast variance. Moreover (5) also captures the possibility that A and B share common information, but A is less efficient at generating forecasts.

Proposition 1: Let trader A be a Blackwell-inferior forecaster. Then (i) $W^A_{eq} = 0$; (ii) $\Pr\{\sup_{0 \le t \le \infty} W^A_t \ge K \cdot W^A_0\} \le 1/K$. Therefore for any fixed $\overline{W}$, the probability that the inferior forecaster's wealth ever reaches $\overline{W}$ (or above) goes to zero as $W^A_0 \to 0$; (iii) $\Pr\{W^A_t > 0 = W^A_{eq}$ for all $t\} \to 0$ as $W^A_0 \to 0$.

Part i of the proposition follows by putting $\sigma^2_A > 0, \sigma^2_B = 0, \theta^2 > 0$ in Figlewski's model and using a limiting argument. Part ii is obtained by invoking theorem 12 in Kushner (1967). Part iii follows from i, ii, and the fact that the states in (0, 1) are all transient, while $W^A = 0$ is absorbing.[2]

IV

Consider the strong form of the efficient-markets hypothesis in which the market price accurately reflects all information available to traders, including private information. Figlewski indicates that his market is efficient when, "given the market price, there should be no further advantage to be derived from having the specific information of the individual market participants" (1978, p. 589). However, this does not seem completely in accord with his formal definitions of X_e, [illegible] Indeed, unless traders share the exact same forecasts, it is [illegible] case that some trader would derive further advantage from the specific information of all traders. The reason for this [illegible] see. Any trader (say A) will only take a nonzero position if he [illegible] to make a capital gain by so doing. That is, $E^A_t(W^A_{t+1} | W^A_t)$ mu[illegible] W^A_t in (2). Now suppose that the market is at Figlewski's [illegible] equilibrium. Then if trader A had access to the specific information of *all* traders, he would know that $E_t(W^A_{t+1} | W^A_t = W^A_{eq})$ does not exceed W^A_t at all but is instead equal to it. That is, on average trader A's gains are exactly offset by his losses when the market is at its long-run equilibrium. Being risk averse, A would clearly choose a zero trading position if he recognized that this was the situation

It is also worth pointing out that in Figlewski's two-trader model, every trader can always infer all private information in the market from the knowledge of his own wealth level, private forecast, total market wealth, and the prevailing market price. All that is required for this to be true is that traders know each other's private forecast variance, degree of risk aversion, and the fact that P_t is determined by equation (3). However this only holds for a two-trader system. With more than two traders, such inferences are generally impossible.

---

[2] By way of contrast, when $W^A_{eq} > 0$, the likelihood that trader A incurs a capital gain when his wealth is W^A_t grows as $W^A_t \to 0$

V

As a final issue, consider whether Figlewski's futures market has a tendency to move in cycles. Figlewski argues that when the market is in its unique long-run equilibrium, the futures price P_t will not only be an unbiased estimator of the spot price P_t^*, but will be minimum variance as well. Moreover, he shows that the market will always have a tendency to move in the direction of the long-run equilibrium. Figlewski is careful to point out that, while it is tempting to call W_{eq}^A a stable equilibrium, he has not "ruled out parameter values that would lead to explosive oscillations around it" (p. 591). Indeed I shall now extend his analysis to show that W_{eq}^A must be unstable (for all parameter values, unless W_{eq}^A is a boundary value in which trader A has either all or none of the wealth in the market).[3]

Let us ask whether trader A's wealth can be expected to move closer to W_{eq}^A with the passage of time. Therefore consider the squared (Euclidean) distance $D(W_t^A)$, $(W_t^A - W_{eq}^A)^2$, between W_t^A and W_{eq}^A. Can we expect $D(W_{t+1}^A)$ to be less than $D(W_t^A)$? The following proposition establishes the existence of an unstable neighborhood (W_l, W_t) of W_{eq}^A. That is, if W_t^A lies inside $(W_l, W_{l'})$ one expects to move further away, and if W_t^A lies outside $(W_l, W_{l'})$ one expects to move closer to W_{eq}^A. Therefore it would not be unreasonable to expect the market to cycle around W_l next period (say) until a large disturbance moves it toward $W_{l'}$, in which case it cycles around $W_{l'}$ for a time. As a result, the market might appear to display a long-run cycling pattern composed of alternating shorter cycles. Interestingly, the long-run cycling pattern will appear to oscillate around W_{eq}^A.

Proposition 2: Let $0 < W_{eq}^A < 1$. Then there exists a neighborhood $(W_l, W_{l'})$ of W_{eq}^A with $W_l > 0$ and $W_{l'} < 1$ such that $D(W_t^A)$ exceeds $E[D(W_{t+1}^A)|W_t^A]$ if W_t^A is outside this neighborhood. If W_t^A is inside the neighborhood, then $D(W_t^A)$ is the strictly smaller of the two. When $W_t^A = W_l$ or $W_t^A = W_{l'}$, then the two are equal.

Proposition 2 follows as a direct consequence of the expression for $E[D(W_{t+1}^A)|W_t^A] - D(W_t^A)$, whose value is easily computed.

The reason for W_{eq}^A not being stable is intimately associated with the discussion in Section IV above. Consider a period when the market begins at the efficient equilibrium. Recall that each trader will typically engage in trade because he (wrongly) expects to incur a capital gain that is strictly positive. Now this trade will almost surely leave him with some change in his wealth level at the end of the period (even though capital losses exactly offset capital gains *on average* at the efficient equilibrium). As a result, the system will move away from the efficient equilibrium. Consequently, this notion of efficient equilibrium has the seeds of instability sown within it. Of course, traders act as if they know neither the location of W^A_{eq} nor the fact that each trader's expected capital gain is zero at W^A_{eq}. Indeed, this seems fairly descriptive of real-world markets. Moreover, if traders did start to reduce the size of their positions as $W^A_t \rightarrow W^A_{eq}$ (as would be necessary for stability), then W^A_{eq} would itself shift in general. The location of W^A_{eq} would stay fixed only if all traders decided to reduce their demands *in the same proportion* (from the original level).

---

[3] As Figlewski notes, any nonboundary value of W_{eq}^A constitutes a transient state, so that his market is generally not stochastically stable in the very long run.

References

Figlewski, Stephen. "Market 'Efficiency' in a Market with Heterogeneous Information." *J.P.E.* 86 (August 1978): 581–97.

———. "Information Diversity and Market Behavior." *J. Finance* 37 (March 1982): 87–102.

Kushner, Harold J. *Stochastic Stability and Control.* New York: Academic Press, 1967.

McGuire, C. B. "Comparisons of Information Structures." In *Decision and Organization: A Volume in Honor of Jacob Marschak*, edited by C. B. McGuire and Roy Radner. Amsterdam: North-Holland, 1972.

Efficient Asset Portfolios and the Theory of Normal Backwardation: A Comment

Alan J. Marcus

Boston University School of Management

In a recent paper, Carter, Rausser, and Schmitz (1983) (hereafter CRS) argue that previous studies of the risk premium on agricultural commodity futures contracts used incorrect proxies for the market index and consequently mismeasured the equilibrium risk premium on those contracts. CRS show that when an alternative market index is used to measure systematic risk, the required risk premium is positive and of economically significant magnitude. This result contradicts that of Dusak (1973) and supports the notion of a generalized Keynesian theory of normal backwardation. The purpose of this Comment is to argue that the market index constructed by CRS is inappropriate and that their empirical results stem directly from the use of this index.

The CRS Model

CRS follow Dusak in employing the capital asset pricing model as the basis for their analysis of equilibrium risk premia. They regress first differences of the logarithms of weekly average futures prices on a market index to derive the measure of systematic risk (beta) for five agricultural commodities. However, they argue that Dusak was incorrect in using the Standard and Poor's 500 Stock Price Index as the proxy for the market index since the S&P 500 excludes agricultural commodities. (Virtually no U.S. farms are traded in equity markets.) Instead, CRS argue that an equally weighted average of the S&P 500 and the Dow Jones commodity futures index would constitute a more appropriate market proxy.

The construction of the market index is the central issue in the CRS methodology. To appreciate its importance, note that the average beta of commodity futures prices against the Dow Jones commodity futures index must, by construction, equal 1.0, while Dusak has shown that the beta of commodities against the S&P 500 is close to

[*Journal of Political Economy*, 1984, vol. 92, no. 1]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0011$01.50

zero. Thus, the beta of a typical commodity futures price against a weighted average of the S&P 500 and commodity index will be an increasing function of the weight assigned to the commodity index. With a zero weight assigned to the commodity index, Dusak found average betas close to zero, but as the weight approaches one, the betas must average to one. With a 50 percent weight, CRS found commodity futures betas that centered approximately at two-thirds. Thus, the estimated beta turns critically on the weight given to commodity futures in the constructed market proxy.

The proper theoretical construct for the market is the value-weighted index of all assets in the economy. If the S&P 500 does underrepresent the share of agriculture in the economy, then CRS are correct in attempting to adjust the index to take greater account of that sector. However, the equal weights given the stock and commodity indexes overadjust the index toward the agricultural sector and thereby bias upward CRS's estimated betas.

A more reasonable weight for the commodity index is roughly one-tenth. This weight can be derived from several sources. First, household sector net wealth in 1977 was $5,845 billion.[1] The value given by CRS for agricultural farm assets is $655 billion, which is 11.2 percent of net wealth. This figure is also consistent with income data. Gross farm income (including government payments) was only 7 percent of national income, and net farm income was 12 percent of corporate profits in 1981 (Council of Economic Advisers 1982).

The justification given by CRS for weighting the stock and commodity indices equally seems to rest on a comparison of stock market value in 1977 ($950 billion) to two measures of the value of the ricultural sector of the economy: total farm assets ($650 bil the value of commodities represented by contracts traded ($1,230 billion). These numbers are, however, misleading.

First, the $1.2 trillion figure for commodity futures cont inappropriate measure of the value of commodities. As Blac has emphasized, futures contracts are zero net aggregate supp sets. Every long position is offset by a short position. The theoretical market index must thus exclude futures contracts. Further, the $1.2 trillion figure for commodity contracts seems to be a serious overestimate of the aggregate value of commodities. In 1977 $1.2 trillion was fully one-fifth of total domestic net wealth. I suspect that this figure represents the sum of all trades in all contracts over the course of the year rather than the more meaningful value of the stock of open interest outstanding at any point in time. Summing the value of the open interest of every agricultural (i.e., grain, oil seed, livestock, food, and fiber) futures contract listed in the *Wall Street Journal* on March 25, 1983, yielded a total of only $19 billion, or less than 2 percent of the $1.2 trillion figure cited. The comparison of farm assets to equity value is also misleading. Ideally, one would like to include all assets in the market index. Although data problems associated with nontraded assets make this impossible, it seems clear that the S&P 500 should proxy for the returns on unincorporated businesses, pension holdings, real estate, residences, and corporate bonds more accurately than would agricultural futures prices. It follows that the weight assigned to agricultural commodities in the market index should be the relative proportion of commodities in total wealth, and not 50 percent.

---

[1] This figure is taken from balance sheets constructed from the Flow of Funds data by the Board of Governors of the Federal Reserve System (1981). Net wealth includes the value of residential structures, consumer durables, land, financial assets, pension fund reserves, and equity in nonincorporated businesses, and is net of private debt such as mortgage and bank loans and consumer credit.

Thus, CRS appear to have assigned far too great a weight to commodities in the construction of their market index. Their procedure casts doubt on the validity of their empirical results: essentially, commodity futures prices were regressed against an index dominated by those very futures prices. Moreover, Bodie and Rosansky (1980) show that almost all commodity futures prices are positively correlated, so that each futures contract will follow the Dow Jones commodity index. It is not surprising that regressions of commodity contract prices on an index comprised so heavily of a commodity price index result in large positive betas. A more reasonable weighting scheme would have given much lower betas, and the conclusions regarding risk premia in commodity futures contracts would be correspondingly reversed.

References

Black, Fischer. "The Pricing of Commodity Contracts." *J. Financial Econ.* 3 (January/March 1976): 167–79.

Board of Governors of the Federal Reserve System, Flow of Funds Section. "Balance Sheets for the U.S. Economy." Mimeographed. 1981.

Bodie, Zvi, and Rosansky, Victor I. "Risk and Return in Commodity Futures." *Financial Analysts J.* 36 (May/June 1980): 27–39.

Carter, Colin A.; Rausser, Gordon C.; and Schmitz, Andrew. "Efficient Assets Portfolios and the Theory of Normal Backwardation." *J.P.E.* 91 (April 1983): 319–31.

Council of Economic Advisers. *Economic Report of the President.* Washington: Government Printing Office, 1982.

Dusak, Katherine. "Futures Trading and Investor Returns: An Investigation of Commodity Market Risk Premiums." *J.P.E.* 81 (November/December 1973): 1387–1406.

Miscellany

Financial Futures Markets and Tabular Standards

Milton Friedman

Hoover Institution

The goal of a monetary system that provides assurance against fluctuations in purchasing power is ancient. One frequently suggested and repeatedly rediscovered proposal is to attain that result by linking the currency unit to a price index. That device was proposed in the nineteenth century by W. Stanley Jevons and by Alfred Marshall, who named it a tabular standard. It has been repeatedly rediscovered. In Marshall's version it required no governmental action except the issuance of a price index number, something which has of course become widely prevalent. Individuals would simply in their own interest be led to make contracts specified not in terms of the nominal number of monetary units—the number of dollars—but in terms of the purchasing power equivalent of the present number of monetary units; that is, the contract would call for paying a given number of
multiplied by the ratio of a price index as of that date to th
which the contract was entered into.

Despite the theoretical attractiveness of this idea and the a
any effective hindrance to its adoption, it has never become
It is a particular version of the generally attractive idea of w
indexation. Yet indexation has been extensive only when infla
been extremely high and variable as in some South American countries and Israel—and has generally fallen into disuse when inflation has been conquered. The only area in which there are any large number of indexed contracts is the area of labor contracts, and even here the use of indexation is by no means dominant and tends to rise and decline as inflation rises and declines.

[*Journal of Political Economy*, 1984, vol. 92, no. 1]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9201-0004$01.50

The recent explosion in financial futures markets offers a very different possible road to the achievement of the equivalent of a tabular standard through private market actions. This possibility is highly speculative—little more than a gleam in one economist's eye.[1]

Its central feature would be the establishment of futures markets in one or more price indexes—strictly parallel to the markets that have developed in stock price indexes. Suppose there were such markets existent or active and covering a considerable range of future dates. Any individual entering into a contract for a future date could hedge himself against inflation uncertainty by use of a futures market. Consider a builder contracting to erect an apartment house under a fixed dollar contract. He could roughly estimate the time distribution of his costs and at convenient intervals enter into a contract to buy the number of dollars he estimates would be required. If inflation rose his costs would go up but he would have an offsetting gain on the futures market. If inflation were less than he had anticipated when he entered into the contract he would have a gain in his construction activity, but that would be matched by a loss in the futures contract.

Similarly, the person who has contracted to buy the building could hedge his risks. Let us assume for a moment that he is buying it with an eye to reselling. He has commited himself to pay a certain number of dollars; he wants to be assured of receiving a corresponding number of dollars. He accordingly sells dollars forward. If the price at which he can sell it goes up beyond his expectations he makes a profit on selling the building, but that is matched by a loss on his futures contract. Conversely, if the price at which he could sell the building plunges, that is matched by a gain on his futures contract.

The great advantage to operating in a futures market by comparison with a tabular standard, which might lead to the wider use of the one than has occurred of the other, is that each party to a contract can act separately. It is not necessary for the seller and the buyer to negotiate a purchasing-power contract; each one can hedge as and when he wishes in the futures market separately. In addition, if futures markets in price indexes did become popular, there would undoubtedly develop a variety of such indexes. There would presumably be futures in the consumer price index, perhaps in the producer's price index, perhaps in components of the one or the other, such as, for example, construction costs or prices of metals and their products. Since the price risks against which buyers and sellers need to hedge are likely to be rather different, the multiplication of such price indexes would offer a greater flexibility to the whole process. The buyer could hedge on one index, the seller on another.

---

[1] Since writing this note I have learned that the Coffee, Sugar, and Cocoa Exchange has an application pending with the Commodity Futures Trading Commission to introduce a futures contract based on the Consumer Price Index.

An incidental advantage of the existence of futures markets in price indexes is to provide market judgments on the elusive concepts of inflation expectations. We have great uncertainty, for example, about how to interpret interest rates on long-term securities. How shall those interest rates be divided between the "expected inflation" and the "real interest rate"? The futures market in the consumer price index would immediately give a reading of what on the average is the expectation in the market.

Book Reviews

Schumpeterian Economics. Edited by Helmut Frisch.
New York: Praeger Publishers, 1982. Pp. xx + 184. $27.95.

This volume presents 10 papers delivered at a symposium held at the Austrian Academy of Sciences in 1980. The list of contributors is a distinguished one: Paul Samuelson, Wolfgang Stolper, Herbert Giersch, Erich Streissler, Arnold Heertje, Martin Bronfenbrenner, Kurt Rothschild, Bruno Frey, Mark Perlman, and Eduard März. Seven of the papers address the central themes and theses of Schumpeter's main works, the *Theory of Economic Development* (1912), *Business Cycles* (1939), and *Capitalism, Socialism and Democracy* (1942). Samuelson's paper is concerned mainly with one specific Schumpeterian proposition (the zero rate of interest in a stationary economy); Perlman's presents an outline of Schumpeter's *History of Economic Analysis* (1954), comparing it with W. C. Mitchell's lectures on *Types of Economic Theory;* and März's gives an account of Schumpeter's brief career as Austrian Minister of Finance in 1919. Six of the authors (Samuelson, Stolper, Giersch, Bronfenbrenner, Rothschild, and Frey) express great admiration for Schumpeter as a social scientist, two (Streissler and Heertje) are strongly critical, and two (Perlman and März) give no general evaluation of him.

The attempt to fix Schumpeter's position in the pantheon of social science began with a special issue of the *Review of Economics and Statistics* (May 1951) shortly after his death, containing papers by 14 of his erstwhile colleagues and students. Since then there have been numerous symposia, books, and papers on him. Yet, withal, his status as a social scientist remains equivocal. The problem is evident in the opening paper of the present volume, by Paul Samuelson. Under the general title of "Schumpeter as an Economic Theorist," Samuelson discusses at length only one technical (and rather unimportant) proposition by Schumpeter, demolishes it, notes in passing some other errors in Schumpeter's economic analysis, and then concludes, without further ado, that nevertheless he was "one of the few most important economic theorists of the twentieth century." This disjunctive style of appraisal is repeated frequently in the volume, as in other writings on Schumpeter, and one remains at a loss to understand the basis for his nomination to the Hall of Fame.

Bruno Frey ("Schumpeter, Political Economist") tells us that he is "thrilled by how many ideas that are now part of modern political economy are already contained in [Schumpeter's] writings," attributing to *Capitalism, Socialism and Democracy* the "essence" of public choice theory (a claim echoed by Rothschild). But Frey goes on to point out that modern public choice theory developed independently of Schumpeter, and it turns out that, instead of

[*Journal of Political Economy*, 1984, vol. 92, no. 1]
© 1984 by The University of Chicago. All rights reserved

crediting it to him, Frey regrets that the theory was *not* inspired by his writings. Schumpeter's effective contribution is reduced to the proposition that politicians and bureaucrats are self-interested, an important point, but one that could as well be attributed to James Mill, or to Machiavelli, or to Juvenal, who raised the central problem of political science in the famous query *Quis custodiet ipsos custodes?*

The work that first brought Schumpeter international repute as an economist was his *Theory of Economic Development*, especially because of his emphasis on innovation and the concept of "creative destruction," his connection of growth with cycles, and his proposition that inflationary credit creation stimulates development. Both the originality and the validity of these ideas are open to question. Streissler points out that the significance of innovation was forcibly argued by Schumpeter's teacher Friedrich von Wieser. Having denied Schumpeter credit for originality in this, he awards it to him for the theory of "forced saving" through credit creation. But, as historians of economics have noted, Henry Thorton had made the same point as early as 1803. If the theory that such credit creation is a *necessary* factor in growth were correct, Schumpeter could be credited for reviving a neglected point, but since it is not he can only be said to have resurrected an old error. Concerning Schumpeter's theory of business cycles, no one has much good to say that goes beyond Stolper's mysterious proposition that economic development and cyclical movement are "synonymous." The theory of the interaction of three differently phased cycles is generally viewed negatively: Stolper calls it a "confusion" and Samuelson a "hocus pocus" that was inspired not by empirical experience but by "the mysteries of harmonic analysis," Schumpeter having a tendency to be entranced by mathematical formulations that he did not fully understand. Giersch opens and closes his paper with praise of Schumpeter as providing ideas that have guided his own studies of economic development. But in between we find a series of rather arbitrary and disconnected assertions about the world economy that are "Schumpeterian" only in that the author calls them so. To describe every innovative change that occurs as "Schumpeterian" is like giving Newton credit for the *existence* of gravity

The mystery of Schumpeter's status as a scholar is deepened by [illegible] as it is by most other writings on him. It can be clarified if we treat [illegible] an economist but as a Laplacean social theorist, the architect of a [illegible] sets out to tackle the Big Questions—where we are; how we have [illegible] here; where we are inevitably going—in the grand style of Augu[illegible] and Karl Marx. The apex of Schumpeter's social thought, if it is [illegible] way, is his *Capitalism, Socialism and Democracy*. The details of this [illegible] been subject to devastating criticism: its dependence on arbitrary and [illegible] definitions; its failure to discuss the economics of socialism in support of the proposition that a nonmarket system can "work"; its contention that the growth of large monopolistic enterprise promotes innovation while simultaneously destroying it through bureaucratic routinization; the lack of empirical support for its main predictions during the 40 years that have elapsed since its publication. But many who admit that Schumpeter's specific contentions cannot stand nevertheless praise him for his "vision." This seems to be Schumpeter's essential appeal. If one is dissatisfied with the current state of economics, one can claim that it would be improved by Schumpeterian vision (Bronfenbrenner, Rothschild, Frey); if one believes that there are laws of history even though none has yet been discovered, one can consider Schumpeter's work as a triumph of faith over skepticism (Stolper); if one dislikes capitalism, one can praise Schumpeter for predicting its demise (Bronfenbrenner); if one rejects objectivity in social science and insists that it should be based on intuitive vision inspired by ideology, one can interpret Schumpeter's reverence for "science" in this way (Rothschild); if one admires anyone who has the *chutzpah* to construct a cosmic theory of society, one admires Schumpeter (Samuelson).

Schumpeter was a complex personality and intellect: an egoist and poseur, yet courteous, kindly, and almost as omniscient as he pretended to be; a romantic and a rationalist; a metaphysical historian who relied heavily on clairvoyance and a hard-working empirical scholar; a practitioner of speculative Grand Sociology and a strong promoter of mathematical modeling and econometrics; a conservative for whom pre-1914 Vienna represented Western civilization at its best and a prophet of what Engels called "scientific socialism." That the papers in this volume do not resolve these contradictions is a merit since any effort to do so would attribute to Schumpeter a coherence of ideas that he did not possess. Unfortunately, that is almost all that can be said in favor of the book. Except for Streissler's insistence that Wieser was the originator of the concept of "Schumpeterian" innovation, it contains nothing of significance that cannot be found in the previous literature. It does, however, shed some light on why Schumpeter has generated such ardent admiration—if one focuses on the votaries rather than the object of their veneration.

Scott Gordon

Indiana University and Queen's University

The Economics of Feasible Socialism. By Alec Nove.
Winchester, Mass.: Allen & Unwin, 1983. Pp xi+244. $29.50 (cloth); $9.95 (paper).

"Socialism" is a term with numerous, complex, and very diverse philosophical, economic, social, and political connotations. As early as 1848, in their "Communist Manifesto," Marx and Engels dismissed a whole series of "socialisms," namely "feudal socialism," "petty bourgeois socialism," "German or true socialism," "conservative or bourgeois socialism," and the "critical-utopian socialism" of Saint-Simon, Fourier, Owen, and others. Marx and Engels promoted, instead, their own "scientific socialism," a theory alleged to be neither a utopian invention nor a discovery of a would-be universal reformer, but an expression in general terms "of actual relations, springing from an existing class struggle, from a historical movement going on under our own eyes" (Marx and Engels 1955, p. 46). In the Marxian schema, socialism is the *predictable* outcome of the decline of a capitalism rent by "inner contradictions." As a political-economic system, socialism comes into being when the proletariat rises through revolution to the position of the ruling class and when it uses this political supremacy "to wrest, by degrees, all capital from the bourgeoisie, to centralize all instruments of production in the hands of the State, . . . and to increase the total of productive forces as rapidly as possible" (Marx and Engels 1955, p. 53). Also, socialism evolves as a system through a *transitional* or *transformation* period extending between the fall of capitalism and the advent of full communism. During that period "despotic inroads in the rights of property" are conducted by the ruling working class in order to revolutionize the mode of production, raise productivity to unknown heights, overcome scarcity, and achieve a nonmonetary, planned, collectivist system in which the state will cease to exist (Marx and Engels 1955, p. 54; see also Engels 1955, pp. 136 ff., 153). Marx's followers have believed that he "laid bare the laws of motion of capitalism," discovered the hidden mainspring of its *modus operandi*, and demonstrated incontrovertibly that the "law of value" (i.e., the play of "elemental" forces of the market determining prices and wages) that dominates capitalism will eventually recede into oblivion under socialism. The main disputes among Marxians have rarely involved these tenets; they have more usually concerned the issues of *how* the proletariat rises to the position of the "ruling class," which specific measures it will enact, and above all at what speed, once it is in power. The basic concepts of "despotic inroads in the rights of property," of "centralization of all instruments in the hands of the State," of the supplanting of market relations with centralized directives (in order to subjugate the "law of value") are embedded in the very foundations of all existing "socialist" countries. The debates on systemic reforms in these countries concern only the extent and the speeds of expropriations, centralization, and limitation of the market forces

Considering Marx's theoretical concepts of socialism and their current applications in the socialist countries, Nove rejects them as "utopian" and as "intolerably inefficient," respectively. He then raises the following questions. "What species of socialism *could* be envisaged? Would such a socialism be free of the defects of the Soviet model and of other 'really existing' variants? Could it operate with reasonable efficiency, and give satisfaction to the citizens in their capacities as consumers and producers?" (p x). In the process of answering these questions, that is, of "fleshing out" a "feasible socialism" as an alternative to societies based on private ownership and on Soviet-type relations, Nove deduces its necessary conditions via a step-by-step critique of Marx's theoretical legacy (pt. 1 of the book), of the actual Soviet experience (pt 2), and of the East European and Chinese adaptations (pt. 3). (It is interesting to note that he does not bother at all with other kinds of socialism) This critique is followed by recommendations concerning strategies of transition from (*a*) a developed or semideveloped capitalist society to socialism, (*b*) a Soviet-type socialism to "something more acceptable", and (*c*) a Third World context to "developmental socialism" (pt. 4). The concluding section (pt. 5) recapitulates and rounds out the necessary conditions for achieving th sought-for "feasible socialism," "without our having to make or plausible or far-fetched assumptions about society, human beings economy" (p. 197)

How are these necessary conditions searched for and found? To while the Marxian prediction about socialism (the socialist "transi period") is open-ended like all prophecies, Nove specifies from the he wishes to set for his "feasible socialism" a definite horizon, this should be achievable "within the lifespan of a generation" (say, 50 yea are not told, however, *how* the socialists come to power and how they there for the life span of a generation. Be that as it may, they are invited to reject as unacceptable the Marxian assumption about the march toward abundance under socialism. Nove notes, instead, that for the next decades, in a worldwide perspective, "the balance of probability is strongly tilted in the direction of the continued existence of scarcity" (p. 17). In conditions of scarcity, the socialists are further advised, "economic calculation, opportunity-cost, choice between alternatives, clearly will have to exist" (p. 20). The "Marxian value theory and its influence on economics of socialism" are either misleading or irrelevant (p. 25). Equally misleading are the ideas about the elimination of the "production of commodities" and its eventual replacement by "production for use" (i.e., of free goods); "one would have commodity production" (p. 30).

Could the market be supplanted under "feasible socialism"? Such a socialism, in which public property will be dominant but not exclusive, could not entertain the Marxian "illusion" that labor could be directly allocated, in a planned fashion, to predetermined tasks. After an excursus on the "irrelevancies" of the labor theory of value, on the controversies about Marx's transformation theory of value into prices, and on Marx and Engels's ideas on the "revolution in the mode of production" translated into the "overemphasis on the primacy of production and the neglect of consumer needs," Nove stresses the importance of *scale* and *complexity* in modern industry. In these conditions, he says, the principles of "freely chosen negotiated contracts (which means autonomy and a species of commodity production)" and of "a system of binding instructions from planning offices," "*must* be combined, *must* coexist" (p. 44). Finally, the "romantic-utopian" Marxian theory that socialism could "overcome" the division of labor typical of capitalism is dismissed as unacceptable for "feasible socialism," since the author has already discarded the labor theory of value. And Nove concludes: "Some hierarchy and subordinates are inseparable in organizing production—and the smaller the role of producers' autonomy and of markets, the greater must be the role of hierarchy and of bureaucracy."

Having pointed out the "realistic" implications for socialism of the continuation of scarcity and of its unavoidable sequels, of scale and complexity of production and of their ramifications, and of management problems and of their attendant issues concerning sanctions and rewards, Nove turns in parts 2 and 3 of his book to the analysis of the "methods, policies and administrative arrangements that exist" in the socialist camp in order to "count the cost of what does not exist" (p. 109). The main arguments cluster now around the questions of how exactly to incorporate micro demands into instructions (pp. 74–75), the "curse" of diseconomies of scale, the problems of innovation and of bureaucratic fragmentation (pp. 75 ff.), and the "necessity of enlarging the decision-making of those on the spot" (p. 90). Nove concludes that "feasible socialism" must be concerned with "how to avoid the overconcentration of power and the overcentralization of management, and the diseconomies of scale, which have been among the major defects of Soviet-type socialism. Inflexibility, inefficiency, human alienation and the unrepresentative nature of authority are part of the same unattractive package" (p. 113). But, except for the assurance that "Marx, were he alive, would most certainly have modified his doctrine in the light of experience" (p. 118), the review of the East European reforms yields few further "necessary conditions" for the achievement of feasible socialism. Moreover, we are left in the dark as to the "cost of what does not exist" (i.e., of the alternatives).

Part 4, on the problems of transition from capitalism, Soviet-type socialism, and Third World morass to feasible socialism, contains a number of interesting asides concerning sociopolitical issues and recommendations. They refer to income distribution, various possible alliances with small business, the danger of counterproductive wage demands at inappropriate moments, the impossibility of democratizing planning without strengthening the "market elements," the problem of "ideology," and some other questions not directly relevant to the main argument line. Indeed, the entire part 4 is neglected in the last section of the book (cf. p. 196), which presents the author's general conclusions.

In the opening line of part 5, Nove proclaims himself ready now "to pass to a picture of possible socialism" (p. 196). However, after a recapitulation of the indicated conditions, he feels the need of injecting some additional assumptions and suggestions for buttressing the search for "feasibility." His first assumption is a political one: "a multiparty democracy, with periodic elections to a parliament" obviously is preferable to "the evils of an imposed one-party system" (p. 197). On the thorny issue of what categories of producer goods and services should exist, he suggests the need of "the coexistence of a wide variety of scale, technique, and consequently of organization, indeed of production relations"—with successive generations free to shift "their predominant opinions and objectives" (p. 200). Further, if any activity can be fruitfully undertaken by any individual "this sets up the presumption of legitimacy" (p. 206). Last but not least, "democratic vote could decide the border between the commercial or market sectors and those where goods and services could be provided free" (p. 208). And Nove concludes, "I hope that, in presenting a feasible form of socialism, I have nowhere slipped into romantic utopianism, have made no far-fetched economic or psychological assumptions Is the resulting picture attractive? Is it desirable? This is very much a matter of opinion" (p. 228).

As we see, Nove does not hide his preferences. Indeed, from the outset he criticizes capitalism (albeit in a cursory manner; see pp 3–5) and then affirms his preference for socialism "with a human face," if one accepts it as "feasible." It is on this basis that he launches confidently into a kind of direct and indirect dialogue with Marxian "fundamentalists" (in power or out of power), as well as with other would-be socialist reformers, in order to show them that they may go astray either if they follow Marx's irrelevancies or if they do not correctly appraise the experience of "existing" socialism Now, one may easily retort that the "fundamentalists" have seized and are maintaining their power on the basis of very definite conceptions concerning the party, the state, and the economy They may correct this or that aspect of their activity; but is it truly "realistic" to expect them to discard the joint foundations of their power—namely, party control, expanding (not contracting) nationalizations priority of centralized directives, and predominance of planners' (not of consumers') preferences?

Nove's "feasible socialism" would involve economic calculation, c production, flexible combinations of planning and markets, r hierarchies in organized production, proper incentives and rew multiparty democracy. Except for Nove's preference for a pr public sector, would not his "feasible socialism" look in many res "mixed" capitalism? Then why bother? In what new respects is the picture attractive"?

Nove is determined not to "slip into romantic utopianism " But "realistic" and not "far-fetched" to assume that the socialists could st power for the "lifespan of a generation" in order to establish their "feasible socialism" within the framework of a multiparty democracy?

Nove's critique of the Marxian legacy, of the Soviet experience, and of the East European and other variants is cogent, to the point, and knowledgeable. It reinforces similar analyses presented by him in his other books But the "necessary conditions" that this critique is supposed to yield for the foundation of his "feasible socialism" are an insufficient base for a theory Indeed, a "feasible" solution derived from necessary conditions (which are supposed to show how and why it is feasible) is no solution at all. One still has to show that these conditions are sufficient as well. What we now have is a presumption built on necessary conditions without knowing if a feasible solution really exists. We do not know whether this solution is utopian (i.e., outside of a feasible set as Nove defines the necessary conditions). If it is not utopian and it is in the set of feasible solutions, then we do not know if this particular solution is the optimal one by some or any criterion. If it is the only solution according to the author's constraints, then it is automatically the *best* and the *worst* simultaneously—by any criterion.

There have been other economists who have tried to deal with the question of what is possible and what is not possible under socialism—Barone, Lerner, and Lange, for example. But they did not presume to make broad "excursions into the future." They simply tried to define narrow issues, approached them technically, and solved them brilliantly, leaving an enduring theoretical heritage.

NICOLAS SPULBER

Indiana University

References

Engels, Frederick. "Socialism: Utopian and Scientific." 1876–78. In *Selected Works*, vol. 2, by Karl Marx and Frederick Engels. Moscow: Foreign Languages Publishing House, 1955.

Marx, Karl, and Engels, Frederick. "The Communist Manifesto." 1848. In *Selected Works*, vol. 1, by Karl Marx and Frederick Engels. Moscow: Foreign Languages Publishing House, 1955.

The Trade-Off between Equality and Efficiency

Edgar K. Browning and William R. Johnson

University of Virginia

This paper uses a 1976 microdata base to estimate the marginal cost of reducing income inequality with a policy that has distributional effects similar to the present tax-transfer system. The analysis uses a simulation methodology in which only the labor supply effects are evaluated; other behavioral effects are ignored. The most striking finding is that marginal cost is quite high even for modest labor supply elasticities. For example, in the benchmark case with a weighted-average economy-wide uncompensated wage elasticity of 0.2 (and compensated elasticity of 0.31), the disposable money income of upper-income quintiles of households is depressed by $9.51 for each dollar increase in the disposable money income of lower-income quintiles. When income equivalent values that take account of the value of leisure are compared, the marginal cost for this case is estimated to be $3.49.

Income redistribution is not a socially costless endeavor because policies required to accomplish it generally produce misallocations of resources. This basic proposition is well known and usually interpreted as implying a trade-off between equality and efficiency. Identifying the magnitude of this trade-off is probably the most important contribution economics can make to the evaluation of distributional policies. However, to be useful, the measure of the trade-off should be a marginal one: What is the cost of slightly more or less inequality than we now have?

Helpful comments were received from Art Snow and Roger Sherman and seminar participants at the University of Virginia and Texas A&M University.

[*Journal of Political Economy*, 1984, vol. 92, no. 2]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0001$01.50

Given the importance of this issue, it is surprising how little attention it has received. Of course, there are many studies that investigate the effects of various transfer and tax policies, and these are certainly related in some way to the marginal trade-off between equality and efficiency.[1] The exact relationship, however, is not obvious since most tax and transfer studies only estimate the total welfare cost of a program, and total welfare costs are not a good indicator of the size of marginal welfare costs, especially in the case of redistributive policies. The optimal income tax literature also deals with this trade-off, but it tends to concentrate on the calculation of optimal marginal tax rates using an assumed social welfare function and hypothetical data, and consequently does not directly estimate the actual trade-off society faces.[2]

In this study, we seek to clarify the trade-off between equality and efficiency by focusing on marginal changes in the current U.S. tax-transfer system. The analysis utilizes a rich microdata base that allows us to develop estimates of the marginal and average tax rates imposed on households by the current mixture of taxes and transfer programs. *Using these data, we simulate the effects of a (small) change in* the distribution of income produced by a demogrant form of redistributive policy, while taking into account the labor supply effects of the changes in marginal and average tax rates. Our most striking conclusion is that the *marginal* trade-off between equality and efficiency is quite severe even when labor supply elasticities are low and despite a modest total welfare cost of the current tax-transfer system. For example, in our benchmark case with an uncompensated labor supply elasticity of 0.2 (and a compensated elasticity of 0.3), each dollar increase in the disposable incomes of the lowest two quintiles requires a reduction in the disposable incomes of the highest three quintiles of more than $9.00.

I. Preliminary Remarks

The nature of the trade-offs between equality and efficiency to be considered can be clarified by using an income distribution frontier. Assume that there are only two individuals, H with high market earnings and L with low market earnings. In figure 1, their market earnings in the absence of any taxes or transfers are OH_1 and OL_1, and this distribution is identified by point A on the income distribution frontier. To generate other points on the frontier, we must specify a mechanism for altering the distribution of income. Suppose the mechanism is a flat-rate tax on income that finances equal lump-sum grants to *H* and *L*. By raising the tax rate from zero to 100 percent and then plotting the resulting combinations of disposable incomes for *H* and *L*, we can generate the entire frontier, *ABCO*.

---

[1] Danziger, Haveman, and Plotnick (1981) survey the literature on transfer programs, while Aaron and Boskin (1980) and Aaron and Pechman (1981) contain several studies on tax policy and numerous references to that literature.

[2] Mirrlees (1971) is the seminal paper. Sheshinski (1972), Sadka (1976), and Stern (1976) are good introductions to the optimal tax literature.

If the market earnings of both individuals are unaffected by the tax and transfer policy, the frontier would have a slope of − 1 throughout; *L*'s disposable income could be increased by $1.00 at a constant marginal cost of $1.00 in disposable income to *H*. Of course, it is possible that if income effects on *H* are strong enough and labor supply rises enough with the tax, more than $1.00 can be transferred to *L* for each $1.00 loss to *H*. In the more important case, however, where the level of market earnings is adversely affected by the redistributive policy, the frontier will take on the general shape shown in figure 1. That is, raising *L*'s disposable income results in an increasing marginal cost to *H*, until we reach point *C*; for further redistribution, the adverse effect on market earnings is so large that the disposable incomes of both individuals fall, and at a tax rate of 100 percent, earnings presumably fall to zero at point *O* on the frontier.

In terms of figure 1, an obvious measure of the trade-off between

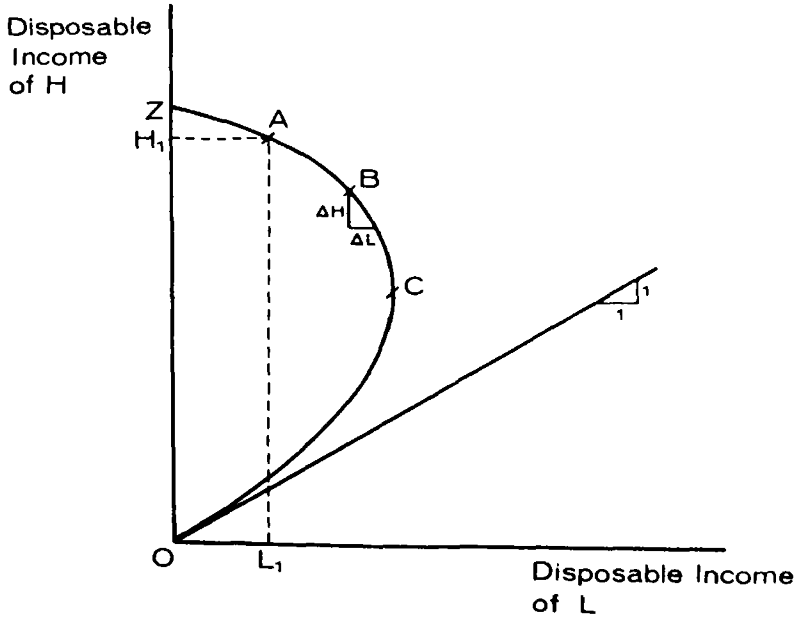


FIG. 1

equality and efficiency is provided by the slope of the income distribution frontier. For example, if the present system of taxes and transfers places us at point B on the frontier, $\Delta H/\Delta L$ measures the marginal cost (to H) of raising L's income. In our simulations, we develop estimates of this trade-off by calculating the change in disposable income for each quintile of households that results from a marginal expansion of the present tax-transfer system. Note, however, that this measure of the trade-off does not identify the welfare effects of changing the income distribution; it only identifies the marginal trade-off in market incomes. The well-being of individuals also depends on their consumption of leisure, and since both individuals in figure 1 will generally be consuming more leisure as we move down the frontier, the trade-off in income-equivalent terms will be smaller than the slope of the frontier. Consider the slope at point B. Since H gains leisure at the same time he sacrifices ΔH in market income, his income-equivalent loss is smaller than ΔH. Conversely, L's gain is greater than ΔL because he gains leisure in addition to ΔL in market income. Besides estimating the marginal trade-off in market incomes ($\Delta H/\Delta L$), we will also develop estimates of the marginal trade-off in income-equivalent terms.

These trade-offs clearly depend on a variety of factors, such as the actual distribution of market earnings, average and marginal tax rates of existing programs, and behavioral responses, all of which will be discussed later. It must be stressed, however, that the trade-off is also likely to depend on the particular policy used to effect a change in the income distribution.[3] In other words, there is no such thing as a unique trade-off between equality and efficiency but instead different trade-offs associated with different policies. Thus, to investigate the issue further, we must specify a policy to be used to alter the distribution of income at the margin.

The policy we will use in our simulations, which we will refer to as a demogrant policy, is a flat-rate tax on labor income (added to preexisting taxes on labor income) that finances equal per person transfers. Although this policy makes transfers to upper-income households, which may appear to bias the results in the direction of finding a large trade-off, it is important to recognize that it is the net effect of the taxes and transfers that is crucial. In fact, this demogrant policy is equivalent to a negative income tax that applies the same marginal tax rate to transfer recipients and taxpayers. Figure 2 illustrates this point. A flat-rate tax collects tax revenues as shown by the schedule

---

[3] For example, Baumol and Fischer (1979) argue that it may be possible to move closer to income equality by using some combination of wage subsidies and taxes (or otherwise manipulating wage rates) rather than by using income taxes and transfers

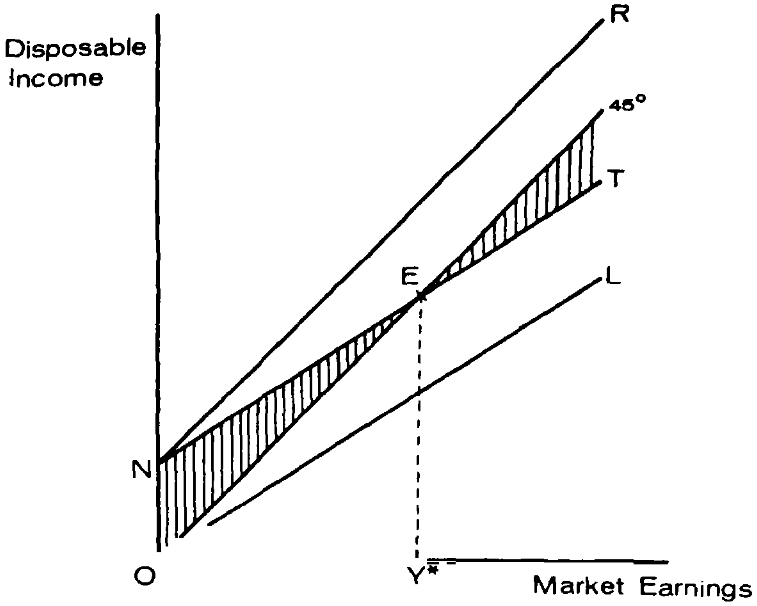


FIG. 2

OL, while equal per person transfers provide benefits (ignoring differences in household size) as shown by the schedule *NR*. The net disposable income at each level of market earnings is shown by the schedule *NET*. Thus, such a demogrant policy is equivalent to tive income tax with a breakeven income of Y^* that makes households with incomes below Y^* and collects taxes from with incomes above that level. For computational reasons, it is to model this policy as overlapping tax and transfer programs economic effects should be the same as for a policy that only or pays out the net sums shown by *NET*.

Some readers have felt that our demogrant policy is so clearly inefficient in redistributing income that we have in effect chosen a straw man. This feeling reflects a confusion concerning the gross and net effects of the tax-transfer system. As line *NET* shows, net transfers are received by low-income families (those with incomes below Y^*), and the greatest transfers are received by the lowest-income families. Of course, it is true that benefits could be concentrated even more on the lowest-income families, but it is far from clear that this would yield a more favorable equality-efficiency trade-off because this would inevitably entail even higher marginal tax rates on transfer recipients who already face high tax rates. Hence, we have chosen to examine a demogrant policy for three reasons. First, we think it is a

TABLE 1

COMPARISON OF DISTRIBUTIONAL EFFECTS

QUINTILE	CHANGES IN DISPOSABLE INCOME WITH DEMOGRANT POLICY ($) (1)	CHANGES IN DISPOSABLE INCOME WITH CURRENT POLICY ($) (2)	DISTRIBUTION OF BENEFITS AND COSTS (%): Demogrant (3)	DISTRIBUTION OF BENEFITS AND COSTS (%): Current Policy (4)
			DISTRIBUTION OF BENEFITS	
1	64	2,581	40	46
2	63	2,388	39	42
3	33	677	20	12
			DISTRIBUTION OF COSTS	
4	−23	−672	14	12
5	−137	−4,963	86	88

reasonably efficient way to redistribute income. Second, it is an easy policy to work with computationally. And third, as we show below, a simple expansion of the current U.S. tax-transfer system would look very much like our demogrant policy.

For our estimates to provide a reasonable guide to the likely magnitudes of the trade-offs produced by marginal changes in actual programs, it is desirable for our demogrant policy to produce distributional results similar to those that would be expected from actual changes in the present system. While it is impossible to demonstrate this directly—since we do not know exactly what marginal changes the political process is likely to generate—it is possible to show that the demogrant policy has distributional effects similar to the entire present tax and transfer system. Consequently, if actual changes simply expand the scale of existing programs our assumed policy should be an accurate replication of this process.

In table 1, column 1 gives the net change in disposable income per household for each quintile that results from an additional 1 percent tax on labor income with the revenues returned as equal per person grants, assuming no behavioral responses. (The data used for these calculations are described more fully below in Sec. III.) Column 2 gives our estimate of the distributional effects of actual taxes and transfers in 1976.[4] A comparison of columns 1 and 2 in table 1 shows that the distributional implications of the demogrant policy are quite similar to those of the present system. In both cases the lowest three quintiles gain at the expense of the top two quintiles. In addition, the percentage distributions of benefits among the lowest three quintiles and costs between the top two quintiles are also similar, as shown in columns 3 and 4. Although the lowest two quintiles receive a slightly smaller share of net benefits under our policy, the difference is small. In addition, we should expect that as the present system is expanded it will be increasingly difficult to concentrate benefits on the lowest-income households: when benefits are small they can be restricted almost totally to the poorest households, but as they become larger a smaller part of each additional dollar transferred can feasibly be granted to the poorest households. Thus, the small deviations in marginal distributional effects of our assumed policy from the average distributional effects of the present system are in the predicted direction.

---

[4] Since total transfers are only 42.6 percent as large as total taxes (because some taxes finance nonredistributive activities like national defense), to impose a balanced budget requirement we assumed that 42.6 percent of each household's total tax burden was used to finance transfers. Combining this "redistributive tax" with the transfer received by each household yields the net change in income due to actual redistributive programs, and the results are given in col. 2.

In addition to having distributional implications similar to the present system, our assumed incremental policy has other advantages. Since taxes are levied on labor income, the major effect of this program should be to influence labor supply decisions, and there is relatively abundant evidence on the magnitudes of the labor supply elasticities required to simulate the effects of the policy (at least as compared with the evidence on saving, e.g.). Per capita transfers are more reasonable than per household transfers since they do not directly encourage household breakup, at least unless income effects have an effect on household composition. (Another possible transfer scheme, distinguishing between adults and children, was examined; in particular, simulations giving children half the adults' transfer revealed no substantial difference from the results presented in the paper.) Finally, we suspect that most economists would agree demogrant policy would probably involve as favorable a tra we are likely to find in practice (because labor is in such supply), so focusing on this policy may give us a good clue minimum cost associated with a movement toward greater equality.

II. A Hypothetical Example

Before turning to the actual simulations, it may prove useful to illustrate with a simple numerical example how various factors interact to determine the size of the marginal trade-off in market incomes. Assume that there are three households, *A*, *B*, and *C*, with labor earnings of $5,000, $15,000, and $25,000, respectively. To take into account the fact that household size increases with income in the United

TABLE 2

ILLUSTRATIVE CALCULATION OF THE TRADE-OFF (All Cols. in $)

Household (1)	Labor Earnings (2)	Change in Earnings (3)	Additional Tax Revenues (4)	Transfer (5)	Change in Disposable Income (6)
A	5,000	-12.50	45	81	23.50
B	15,000	-37.50	135	162	-10.50
C	25,000	-62.50	225	162	-125.50
Total	45,000	-112.50	405	405	-112.50

States, let household *A* contain 1.5 persons, while households *B* and *C* each contain three persons. All households are currently subject to an effective marginal tax rate of 40 percent. To evaluate the effect on labor supply and earnings (it is assumed that the market wage rate is constant), we assume that the elasticity of labor supply relative to the net marginal wage rate is 0.15. To simplify this example, we also assume income effects on labor supply are negligible, but income effects will be taken into account in the actual simulations.

Given these assumptions, table 2 shows how we can calculate the effects of an additional 1 percent tax on labor income with the net proceeds used to finance equal per capita transfers. The 1 percent tax reduces the net marginal wage rate from 60 to 59 percent of the market wage rate, a reduction of 1.66 percent. With an elasticity of 0.15, labor supply and earnings decline by (0.15)(0.0166), or by 0.25 percent. The absolute reduction in earnings for each household is shown in column 3. The revenues generated by the 1 percent supplemental tax alone equal $450, but because of the $112.50 reduction in earnings, the revenues from the initial 40 percent rate decline by $45. Thus, the net increase in tax revenues is $405, and its distribution among households is shown in column 4. Divided on a per capita basis, the per capita transfer is $54, and the transfer for each household is shown in column 5.

Column 6 then gives the change in disposable market income for each household—the sum of the changes shown in columns 3, 4, and 5. The $23.50 increase for household *A* is associated with a reduction of $136 for households *B* and *C*: each dollar increase for *A* reduces the combined incomes of *B* and *C* by about $6.00, or a 6 : 1 marginal trade-off in market incomes. Although based on hypothetical data, this example illustrates how modest labor supply responses can translate into a surprisingly large trade-off.

Intuitively, the explanation of why a large trade-off results is that we are starting from an already highly distorted position with relatively high marginal tax rates. A given change in marginal tax rates then represents a greater percentage change in the after-tax wage rate. Moreover, any decline in labor supply reduces tax revenues under initial tax policies more when marginal tax rates are high. Finally, it is important to recognize that marginal tax rates rise much more than proportionately to the amount of income redistributed. Using our actual data on labor incomes, we calculate that the policy described (assuming no behavioral response) increases marginal tax rates by 4.5 percentage points for each 1 percent of labor income redistributed from the top two to the bottom three quintiles. Since marginal tax rates are the source of the adverse incentive effects, it is not surprising that a small change in income distribution can produce substantial effects.[5]

III. Data

We developed the data used in this study as part of a project that investigated the distribution of tax burdens by income classes (Browning and Johnson 1979). The starting point was the Current Population Survey (CPS) for March 1975, which collected income and demographic data for a representative sample of 47,414 U.S. households. The data were subsequently adjusted for underreporting and updated to correspond to 1976 levels (see U.S. Congress 1977). Several government transfers that were not already recorded by the CPS were imputed to households, including the major in-kind transfers received by households.[6] In addition, two types of factor income, the net rental value of owner-occupied housing and accrued capital gains on stock, were also imputed.[7] Finally, a competitive tax incidence model was used to allocate the burdens of all major taxes to households

A more detailed description of the data base, including explanations of the methods of allocating taxes and transfers and [illegible] various types of income to households, is contained in Brow [illegible] Johnson (1979). The final data set is, we believe, an unusu [illegible] plete representation of the income received by households [illegible] effects of government tax and transfer programs in 1976

Table 3 gives the information that is most relevant to the [illegible] study in summary form. Households have been ranked by before [illegible] income and partitioned into quintiles because our results will be presented on a quintile basis. The figures in the table give the average value per household for each quintile.

---

[5] A further elaboration of these points is contained in Browning (1978)

[6] These imputations were performed by Mathematica Policy Research for the Congressional Budget Office.

[7] The data on net rental values by income class were based on the Brookings Institution's MERGE file.

TABLE 3

RELEVANT CHARACTERISTICS BY QUINTILE (per Household Values)

	QUINTILE				
CHARACTERISTIC	1	2	3	4	5
1. Before-tax, before-transfer income	$2,438	$7,845	$14,357	$21,715	$46,573
2. Taxes	689	2,181	4,142	6,547	17,484
3. Transfers	2,874	3,317	2,441	2,117	2,485
4. After-tax, after-transfer income	4,623	8,981	12,656	17,285	31,574
5. Percentage share of col. 4	6.15	11.96	16.85	23.01	42.03
6. Gross labor income	1,781	6,589	12,691	19,255	31,177
7. Percentage shares of labor income	2.40	9.22	17.75	26.93	43.61
8. Persons per household	1.6	2.5	3.1	3.3	3.4
9. Percentage of population	11.5	18.0	22.3	23.7	24.5
10. Marginal tax rate	.544	.471	.408	.388	.446
11. Average tax rate	−.411	−.102	.102	.186	.305

The tax used in our simulation is a flat-rate tax on labor income, so the distribution of gross labor income in row 6 indicates as a first approximation how the tax will be distributed among quintiles (ignoring the changes in labor income produced by the policy). The net receipts of the tax will be returned as equal per capita transfers, so transfers per household will vary with the average size of households. Because family size rises with income, upper-income households will receive larger transfers than lower-income households. On balance, however, the tax-plus-transfer redistributes income downward. If there were no change in labor income, the lowest three quintiles would gain because their share of total population (row 9) is greater than their share of the tax base (row 7).

The estimates of marginal and average tax rates are especially important to this study because we model changes in labor supply as a function of changes in marginal and average rates (as described in the next section). The effective marginal tax rate for each household should indicate how much disposable income changes for a small change in before-tax earnings, and this depends on the combined effect of all taxes and transfers. (The implicit marginal tax rates of transfer programs are particularly important for low-income households.) Our general procedure in estimating effective marginal tax rates was to rank households on the basis of their factor earnings and then use the average reduction in transfers and the increase in taxes as we move from one income level to another (in $1,000 increments) as the basis for calculating the effective marginal tax rates.

This general procedure was applied to subsets of the entire population of households. Different-sized households were considered separately, and households with heads over 65 years of age were omitted from the samples. Because both personal income taxes and most transfers vary by family size and earnings, separate estimation for different household sizes is important. We eliminated the elderly from the samples to remove the bias introduced by age-related (as distinct from income-tested) transfers. When households are ranked on the basis of their factor earnings, elderly families are disproportionately represented at the bottom. What we then find is that social security and other transfers targeted on the elderly fall as we move up the earnings distribution. This, however, is not the result of marginal tax rates implicit in these programs but occurs because there are fewer elderly households higher in the earnings distribution. Thus, including the elderly would misleadingly imply that social security benefits contributed to high marginal tax rates for low-income households.

Each household in the sample is assigned a marginal tax rate using the procedure described.[8] Row 10 of table 3 gives the weighted average values of the marginal tax rates for the households in each quintile. As expected on the basis of previous research, these rates are generally fairly high, especially for low-income households that frequently receive income-tested transfers.

Our procedure for estimating marginal tax rates is not without defects. In particular, we interpret the estimated rates [illegible] to any changes in gross labor income, yet the estimat [illegible] weighted averages of the marginal tax rates that apply [illegible] capital and labor incomes. Insofar as the true marginal [illegible] applies to capital income is greater than the rate that ap [illegible] income, our estimates will be biased upward.[9]

On the other hand, there is at least one bias operating in th [illegible] site direction that can best be explained with an example. Suppose a certain type of income were taxed at a rate of 100 percent. Such a tax would presumably raise no revenue, and our procedure based on how actual taxes vary with income would estimate a zero marginal tax rate—lower than the true value. In general, our data already reflect behavioral changes made to minimize the burden of taxes and so to some extent tend to understate the effective marginal rate of taxation.[10]

---

[8] The elderly are included in the final data used in the simulations. They are assigned marginal tax rates corresponding to their factor earnings using the procedure described in the previous paragraph.

[9] Although some taxes can be reasonably assumed to fall exclusively on capital or labor income, it is not clear how to treat personal income taxes or transfer payments. Thus, we did not attempt to estimate separate marginal tax rates for capital and labor income.

For these reasons, our estimates of marginal tax rates on labor income may be biased, but the direction of the bias is not clear. In addition, it should be recalled that changes in the tax system since 1976 have operated to increase tax rates applying to labor income. Bracket creep due to inflation has increased marginal rates under the federal income tax,[11] and the 1977 changes in the social security payroll tax have similarly increased effective marginal rates.

Average tax rates are given in row 11 of table 3. For purposes of this study, a household's average tax rate is defined as (total tax − total transfer) ÷ (before-tax, after-transfer income).[12] This rate, or, more precisely, changes in it, is used to indicate the direction of the income effect of taxes and transfers together on labor supply. Note that the average tax rate defined in this way can vary from −1 to +1 and will be zero for a household that receives transfers exactly equal to the taxes it pays.

For purposes of indicating the income effect of the present system on labor supply, this definition of the average tax rate suffers from the neglect of benefits of nontransfer government expenditures like defense and highways. For the community as a whole, we estimate a positive average tax rate of 16.77 percent, but it can be argued that the relevant rate is really zero since all tax revenues presumably finance expenditures that benefit someone. It is not clear, however, how we should distribute the benefits of nontransfer expenditures among households. More important, for purposes of simulating the effect of a *change* in taxes and transfers while holding nontransfer expenditures fixed, we feel the proposed measure is adequate to indicate the direction of the income effect on labor supply.

---

[10] In a recent paper, Barro and Sahasakul (1983) make this point and demonstrate that our method of computing marginal tax rates is likely to understate the effective marginal tax rate relevant for labor supply responses. In this respect, our estimates of the efficiency cost of greater redistribution are biased downward

[11] For four-person families with median income, the marginal tax rate (applying to taxable income) of the federal individual income tax rose from 22 percent in 1975 to 24 percent in 1980. For families with one-half and double the median income, the increase was from 17 to 18 percent and from 32 to 43 percent, respectively (*Economic Report of the President* 1982, table 5-4).

[12] Most tax incidence studies define the average tax rate as the total tax divided by before-tax, after-transfer income. For a study that examines only taxes this is perhaps appropriate, but to use the average tax rate as an indication of the direction of income effect on labor supply for simultaneous changes in taxes and transfers, it must be defined differently

IV. Modeling Labor Supply Responses

The most important ingredient in our simulations is the presumed form of the household's labor supply response to tax and transfer changes. Since we use a way of approximating these changes that differs from most previous work, a justification of our choice is in order. Hours of work with a given set of tax and transfer policies (Y_1) are given by

$$Y_1 = (1 - MTR^{\beta})(1 + \alpha ATR) \cdot Y_0, \tag{1}$$

where Y_1 is labor supply, MTR and ATR are the marginal and average tax rates, and Y_0 is labor supply with no taxes or transfers at all. The parameters α and β can be changed to describe a spectrum of possible labor supply responses. In particular, β, which is restricted to be nonnegative, is related to the substitution effect; the smaller the value of β, the more labor supply responds to changes in marginal tax rates. The parameter α, also presumed to be nonnegative if leisure is a normal good, captures the income effect; the greater the value of α, the more labor supply rises when average tax rates reduce income.

There are three advantages to our labor supply specification. First, our formulation is agnostic about the determinants of Y_0 (which is presumably affected by preferences, wage rates, and nonearned income) and only restricts the response of labor supply to changes in taxes and transfers. We are interested in predicting changes in the pattern of labor supply (movements away from point B in fig. 1) rather than explaining why we happen to be at point B initially. Second, equation (1) is compatible with labor supply curves that are upward sloping at low net wage rates and backward bendi
net wage rates. This property is not shared by many con
utility functions such as the CES function or by consta
labor supply functions. Third, our formulation requires
supply fall toward zero as marginal tax rates approach 10
characteristic important because of theoretical considera

Equation (1) implies that labor supply elasticities depend o
level of average and marginal tax rates and will therefore differ across households. Table 4 provides some idea of the relationship between the parameters α and β and the elasticities they imply at the marginal and average tax rates currently faced by U.S. households. For each household we can calculate a compensated and uncompensated elasticity implied by α, β, and existing tax rates.[13] Weighting these elas-

---

[13] Our compensated elasticity is obtained by changing marginal tax rates but keeping average tax rates constant, hence "compensating" for the income lost by the higher marginal tax rate. Our uncompensated elasticity is computed by allowing both average and marginal tax rates to move by the same amount, similar to a proportional change in wages

TABLE 4

IMPLIED LABOR SUPPLY ELASTICITIES FOR VARIOUS VALUES OF α, β

	PARAMETERS			QUINTILE					
CASE	α	β	ELASTICITY	1	2	3	4	5	OVERALL
A	2	3 25	Uncompensated	435	263	.169	138	.213	.204
			Compensated	.513	364	.283	.255	.318	312
B	.1	2.5	Uncompensated	.600	.464	382	355	.423	.412
			Compensated	639	.515	.439	.414	.475	.467
C	3	2.5	Uncompensated	.523	.366	273	.243	.323	.308
			Compensated	.639	.515	.439	.414	.476	.467
D	1	4.0	Uncompensated	.377	205	.122	.094	.155	.152
			Compensated	.417	256	.179	.154	.209	.207
E	3	4 0	Uncompensated	299	.106	012	−.018	.054	.047
			Compensated	.417	.256	.179	.154	209	.207

ticities by hours currently supplied to the market yields average elasticities for each income quintile and for the entire population of households. Table 4 shows that elasticities are greater for lowest- and highest-income households facing relatively high marginal tax rates. The parameter β strongly affects the compensated elasticities while α, the income effect, determines the difference between the compensated and uncompensated elasticities. By suitably adjusting these parameters, therefore, we can generate a wide range of values of compensated and uncompensated elasticities.

Table 4 and our subsequent simulations embody the simplifying notion that each household, regardless of demographic composition, behaves according to the α and β for the case in question and that any changes in labor supply earn income at the household's average wage rate. Thus, each household is treated as a single individual for labor supply purposes and all individuals are bound by the same α and β. We do not think this simplification biases the results unduly; at least, if it does, the direction of the bias is not clear a priori.[14]

We consider case *A* in table 4 to be the most plausible assumption for α and β, for the reasons described below. The implications of this "benchmark" combination of α and β are given in table 5, which shows labor supply relative to the no-tax case for various combinations of *ATR* and *MTR*. By reading down the columns in table 5 we see the effect of changing marginal tax rates (the "pure" substitution effect), and reading across the rows identifies the impact of income

---

[14] While high-income households have more prime age married men (low elasticity), they also have more secondary workers (high elasticity). Supply elasticities of both taxpayers and transfer recipients are important.

TABLE 5

Labor Supply Relative to No-Tax Case (α = 2. β = 3.25)

	ATR										
MTR	– 5	–.4	–.3	–.2	– 1	0	1	.2	.3	.4	.5
0	.900	920	.940	960	.980	1 0	1.02	1 04	1 06	1.08	1.1
.1	.899	919	939	959	.979	999	1 019	1.039	1.059	1.079	1.099
.2	895	915	935	.955	975	995	1 015	1.034	1.054	1.074	1.094
.3	.882	.902	921	941	960	.980	1 000	1 019	1.039	1 058	1 078
.4	854	.873	892	911	930	949	968	987	1.006	1.025	1.044
.5	805	823	841	859	.877	895	.913	931	949	.966	.984
6	729	.745	.761	777	794	809	826	.842	.858	875	.891
.7	618	631	645	.659	.673	.686	.700	.714	.727	.741	.755
.8	464	475	485	495	505	516	526	.536	547	.557	.567
9	261	267	273	278	.284	290	.296	.302	.307	.313	.319
1 0	0	0	0	0	0	0	0	0	0	0	0

effects. Finally, following the diagonal beginning with $MTR = ATR = 0$, we can plot the uncompensated labor supply curve.

As table 5 suggests, the effect of the existing tax and transfer system on labor supply has been relatively modest, if these parameter values are valid. The median household, with an $ATR = 0.1$ and $MTR = 0.4$, has reduced labor supply about 3 percent as a result of the tax-transfer system, while low-income households ($ATR = -0.41$, $MTR = 0.54$) have probably reduced labor supply by 20 percent or so. Taking all households together, we estimate a total reduction of labor supply caused by the tax-transfer system of 5.2 percent. This compares with the estimate of Danziger, Haveman, and Plotnick (1981) that transfer programs alone (excluding the taxes that finance them) have reduced overall labor supply by 4.8 percent. Lampman (1981) has also estimated the reduction of labor supply due to social welfare programs and the taxes that finance them at 7 percent. Hence, our benchmark parameters yield estimates of the labor supply effect of the current tax-transfer system that are similar to other estimates in the literature.

Table 5 also implies a backward-bending labor supply curve; reading down the diagonal along which $ATR = MTR$, we see that labor supply rises until tax rates reach 30 percent and does not fall below the no-tax position until taxes are 50 percent.

The most important justification of the benchmark parameters would come from comparison with econometric evidence of labor supply response. Unfortunately, that voluminous literature is not easily summarized, nor is there a consensus on labor supply elasticities. Recent critical surveys of the labor supply literature yield some bounds on elasticity estimates. Keeley (1981), surveying the nonexperimental literature, gives a mean estimate of uncompensated and compensated elasticities for married men of −0.11 and 0.10, respectively, with 1.28 and 1.05 as the corresponding figures for married women. Since women work around 30 percent of all hours, a weighted overall average would be 0.3 and 0.37, respectively.[15] Keeley's survey of the experimental literature gives a mean overall compensated elasticity of 0.16.

Moffitt and Kehrer's (1981) survey offers similar averages with generally lower elasticities found in the experimental results. On the other hand, Bishop (1980) contends that the true long-term labor supply response implied by the experiments is larger than conventionally estimated. He finds the overall impact of the Seattle-Denver experiment (the most recent and best executed experiment) as reducing labor supply by about 16 percent. This also accords fairly well with our benchmark case, since table 5 shows the effect (starting from no taxes) of *MTR* = 0.5 and *ATR* = −0.3—which we believe is representative for the average participant in the experiments—as a 16 percent reduction in labor supply. The actual experiments, of course, did not start from the no-tax position, so our benchmark implies a more limited response than in the Seattle-Denver experiment.

---

[15] Other surveys by Fullerton (1982), Heckman, Killingsworth, and MaCurdy (1981), and Killingsworth (1981) are in general accord with these nonexperimental findings.

Some recent labor supply studies have found much higher substitution elasticities than the research summarized above. Hausman (1981) estimates an inelastic uncompensated labor supply response for married men, but substantial compensated effects; using his estimates, we compute an uncompensated elasticity of about 0.3 and a substitution elasticity of about 0.6 for all workers. MaCurdy (1980), using a life-cycle model, finds uncompensated and compensated elasticities of 0.7 and 1.47 for men, respectively. Both of these estimates seem implausibly large to us but suggest that the benchmark case is not likely to overstate labor supply responses.

In short, the benchmark case *A* accords roughly with the bulk of the econometric evidence of labor supply elasticities, evidence of the effect of the Seattle-Denver Income Maintenance experiments, and estimates of the effect of the current tax-transfer system on labor supply. Cases *B* and *C*, with their implied overall economy-wide uncompensated labor supply elasticities of 0.41 and 0.31 (and compensated elasticities of 0.47), seem to us to be about as large as can be reasonably defended. At the other extreme, cases *D* and *E*, with their implied overall uncompensated elasticities of 0.15 and 0.05 (and compensated elasticities of 0.21), would appear to be about as l plausible given available evidence. Although we find the be case the most congenial set of parameters, we offer simulatio for all five cases, hoping that they will span the range of pos most readers find acceptable.

V. Simulation Results

Before discussing the results of the simulations, we should clarify several assumptions that were employed in the computations.

1. Wage rates are assumed to be constant. Thus, gross market earnings vary in proportion to the change in the quantity of labor supplied.

2. Capital income, in the aggregate and for each household, is assumed to be unaffected by the changes in tax and transfer policies being investigated.

3. The marginal tax rate imposed on each household by the current tax-transfer system is assumed to be constant. The change in net

revenues collected from (or dispersed to) each household by existing policies when market earnings change is calculated using this marginal tax rate. This procedure ignores possible nonlinearities in budget constraints—which we believe is plausible for the small changes in taxes and transfers and, hence, small changes in labor supply that we emphasize.

4. For purposes of computing the change in the labor supply for each household, the change in the average tax rate is estimated at current levels of market earnings. In a nonproportional tax system, the average tax rate as conventionally defined is endogenous since it depends on the amount of labor supply chosen. To apply our formula relating labor supply to tax rates, however, this rate must be exogenously given. Our procedure effectively uses the vertical displacement in the budget constraint from the initial equilibrium as an exogenous measure of the change in the average tax rate.

5. The supplemental tax on labor income and per capita transfers are subject to a balanced budget requirement. In practice, satisfying the balanced budget requirement necessitated repeated iterations since the size of the transfer that could be financed by a 1 percent tax depends on the labor supply response, and the labor supply response depends in part on the size of the transfer (through its effect on the average tax rate). In each simulation, therefore, we varied the per capita transfer until the balanced budget requirement was approximately satisfied.

More precisely, we estimated the labor supply effects of a 1 percent additional tax on labor income (and associated transfers) as follows. Letting MTR_1, ATR_1, and Y_1 stand for current marginal and average tax rates and labor supply, we have, using equation (1):

$$Y_2 = \frac{[1 - (MTR_1 + .01)^{\beta}](1 + \alpha ATR_2) \cdot Y_1}{(1 - MTR_1^{\beta})(1 + \alpha ATR_1)}, \qquad (2)$$

where Y_2 is labor supply with the incremental policy and ATR_2 is the average tax rate incorporating the additional transfers as described above. Equation (2) expresses labor supply with our demogrant policy as a function of known quantities.

The Money Income Trade-Off

Table 6 presents our results for a marginal change in the tax-transfer system using a 1 percent tax on gross labor income to finance equal per capita transfers. For each combination of parameter values, we give the changes in disposable money income per household for each quintile. For purposes of comparison, the last row illustrates how

TABLE 6

EFFECTS OF 1 PERCENTAGE POINT INCREASE IN TAX RATE

CASE	PARAMETERS		UNCOMPENSATED AND COMPENSATED ELASTICITIES		CHANGE IN INCOME PER HOUSEHOLD					TRADE-OFF RATES	
	α	β			Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	Losses/ Gains	Losses of Top 4/ Gain of Bottom
A	2	3.25	20	31	$36.49	7.54	−43.15	−108.17	−267.49	9.51	11.27
B	1	2.5	41	47	22.85	−19.60	−84.50	−164.41	−346.85	26.93	26.93
C	3	2.5	31	47	25.23	−15.13	−76.46	−150.65	−324.95	22.48	22.48
D	1	4.0	15	21	44.56	23.83	−20.35	−80.71	−229.59	4.83	6.88
E	3	4.0	05	21	45.73	25.60	−16.11	−71.63	−211.82	4.20	5.99
F	0	∞	0	0	64.48	62.69	32.54	−22.82	−136.89	1	1

money incomes are affected if labor supply does not change at all. Note that this assumption of no labor supply response yields estimates of gains and losses that are quite far off the mark. Even for our most inelastic labor supply response, case E with an uncompensated elasticity of 0.05, the gain to the lowest two quintiles together is only half as large as the gain estimated when labor supply responses are ignored.

Our primary concern here is with the marginal trade-off implied by a movement toward greater income equality. In a many-person world, however, it is not clear how to measure this trade-off most meaningfully. Two measures are given in table 6. The first, in the next-to-last column, gives the ratio of combined income losses for the quintiles that have their disposable incomes reduced to the combined income gains for the quintiles that gain.[16] For our benchmark case A, this ratio is 9.51, implying that at the margin each dollar of disposable income for the lower two quintiles costs the upper three quintiles \$9.51. (A convenient way to think of this ratio is as a measure of the marginal cost of changing the disposable income of lower-income households.)

This measure of the trade-off is somewhat inaccurate because the identities of the quintiles that gain and lose depend on the labor supply response. In case A, for example, the ratio gives the loss of the top three quintiles relative to the gain of the bottom two, while in case B it gives the loss of the top four quintiles relative to the gain of the lowest quintile. An alternative way to measure the trade-off avoids this problem by aggregating the change for the top four quintiles in all cases and comparing this to the gain for the bottom quintile. This ratio is given in the last column of table 6 and may be more appropriate if our goal is primarily to assist households in the lowest quintile.

As will be noted, these measures of the trade-off are quite sensitive to the assumed labor supply response, varying from 4.20 (5.99) for case E to 26.93 for case B. One surprising finding, however, is that the trade-off is not very sensitive to the value of α. Note that keeping β constant but increasing α from 0.1 to 0.3 (cases B and C, or D and E) has only a small effect on the trade-off. We have experimented with values of α up to 1.0 with similar results. Increasing α—which implies a larger income effect on labor supply—reduces the benefit to households that gain (they earn less with a larger income effect) but also reduces the loss to households that lose (they earn more with a larger income effect), so the net effect on the ratio is theoretically indeterminate. Given our formulation, these offsetting effects approximately cancel out insofar as the trade-off is concerned, so the trade-off depends mainly on compensated labor supply elasticities. This conclusion is quite important, particularly if recent research (Hausman 1981) suggesting that compensated labor supply responses are much larger than uncompensated responses (and larger than our value for case *B*) is supported by further work.

---

[16] Note that this measure ignores intraquintile redistribution. Within a quintile, larger-sized households gain more (or lose less) from the policy than smaller-sized households.

Although our major emphasis is on the marginal trade-off, we have also considered a large change in the tax-transfer system designed to increase the money income of the lowest quintile as much as possible. (This corresponds to point *C* in fig. 1, with the aggregate income of the lowest quintile measured horizontally.) This maximin outcome, first proposed by Rawls (1971), presumably defines the outer limit to which anyone would be willing to push income redistributive policies.[17] Although few are likely to be as egalitarian as Rawls, identifying how close we have already come to maximizing the income of the lowest quintile is of some interest. (Actually, Rawls proposed maximizing the position of the worst-off individual, but in a world with considerable mobility over time it is impossible to identify this individual so we have chosen to focus on the lowest quintile of households.)

Table 7 gives the results of these simulations. For our benchmark case *A*, a supplemental tax rate of 19 percent on labor income financing per capita transfers maximizes the income of the lowest quintile, producing an increase in average income of $376 per year per household, an increase of 8.1 percent. Thus, for this case, the current tax-transfer system has already increased the disposable income of the lowest quintile to 92.5 percent of its maximum Achieving an additional gain of $376 results in a loss for the quintiles that is 30 times as great. (In these simulations, the quintiles lose in all five cases, so there is only one loss/gain consider.) For the other four cases, there is the expected v results, though the range of variation is not as pronounced marginal change simulated earlier.

Since we are here investigating a very large discrete change in the tax-transfer system, we believe less significance should be attached to these estimates than to those for a marginal change. Taken at face value, however, the estimates in table 7 seem to suggest that the maximum extent to which disposable incomes in the lowest quintiles can be increased is rather limited, at least when using a tax-transfer policy of the type considered.

---

[17] Of course, "envy" could push the optimal point beyond point *C*

TABLE 7

MAXIMUM INCREASE IN LOWEST QUINTILE'S INCOME

CASE	PARAMETERS		ELASTICITIES		SUPPLEMENTAL TAX RATE	GAIN TO LOWEST QUINTILE		LOSSES / GAINS
	α	β				$	%	
A	.2	3.25	.20	31	19	376	8.1	30.3
B	.1	2.5	41	47	14	156	3.4	66.2
C	.3	2.5	31	47	15	200	4.3	52.2
D	.1	4.0	15	21	22	532	11.5	22.1
E	.3	4.0	.05	21	23	597	12.9	19.4

Welfare Trade-Off

this point, our measures of gains and losses have been in terms of nges in disposable money incomes. These measures, although of e interest, are not good estimates of the changes in well-being ause they ignore the value of additional leisure time to house- lds. For the marginal change considered in table 6, we now convert e gains and losses into estimates of the income-equivalent gains and sses that value leisure. As expected, this will reduce the trade-off ratios since, when the value of leisure is included, the gains to lower-income households will be larger while the losses to upper-income households will be smaller. Nonetheless, the ratio will still exceed unity because of the additional welfare cost that results from a higher marginal tax rate on labor income.

Figure 3 can be used to illustrate how the marginal welfare cost

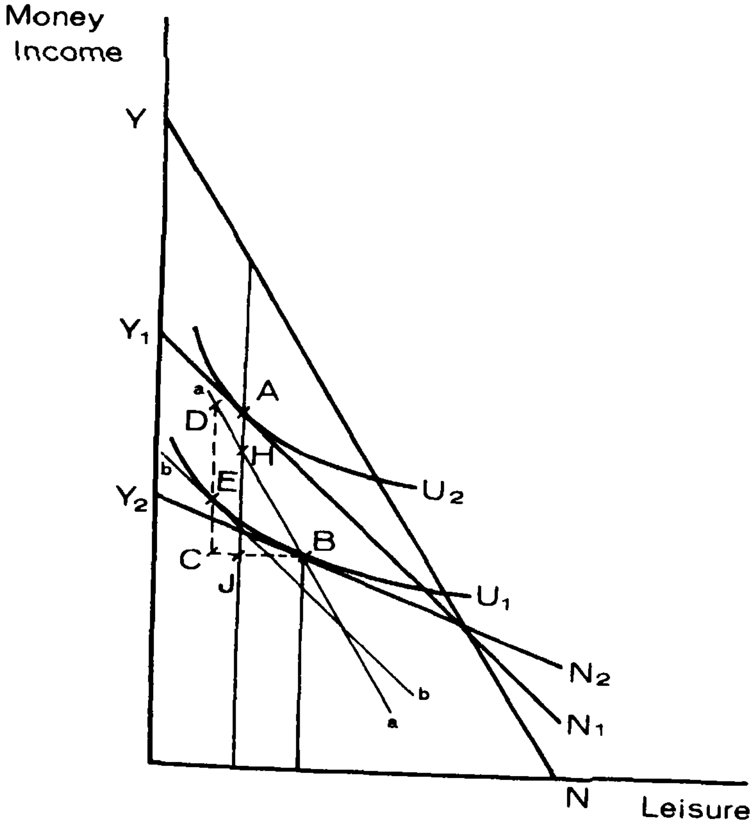


Fig. 3

enters into the calculation of the income-equivalent loss to a taxpayer. (A symmetrical analysis applies to net transfer recipients, whose additional transfers are larger than additional taxes.) The budget constraint in the absence of any taxes and transfers is *YN*, but the current tax-transfer system confronts the taxpayer with budget constraint Y_1N_1, so the initial equilibrium is at point *A*. An additional 1 percent tax on labor income combined with a lump-sum transfer shifts the budget constraint to Y_2N_2—drawn greatly exaggerated for clarity. The final equilibrium is at point *B*, and the reduction in money income is shown by the distance *AJ*.

The income-equivalent loss in well-being is the sum of the additional tax revenue collected from the taxpayer and the additional welfare cost due to the added labor supply distortion from the higher marginal tax rate. Drawing line *aa* through point *B* and parallel to *YN*, we see that the additional tax revenue is equal to *AH*. To identify the marginal welfare cost, we assume that a lump-sum tax is added to the initial tax-transfer system (Y_1N_1) such that the individual can just attain U_1: this is shown by line *bb*, which is parallel to Y_1N_1. This identifies the compensated change in labor supply from the 1 percentage point increase in the marginal tax rate, *CB*, that must be used to calculate the marginal welfare cost. The marginal welfare cost equals the difference between the market valuation of this compensated labor supply change, *DC*, and the individual's valuation of this quantity of leisure, *EC*; thus, *DE* is the marginal welfare cost. An alternative way to see this is to note that at point *E* the taxpayer is bearing a tax burden that is greater than the additional revenue (*AH*) generated by the policy by the distance *DE*. Thus, *DE* is a loss in addition to the direct revenue cost of *AH*. The important point is that the compensated (not the actual or uncompensated) change in labor supply must be used to estimate the marginal welfare cost.

In our simulations, the income-equivalent loss (or gain) was estimated in the following way. First, the compensated change in labor supply is estimated by using equation (2) to calculate the labor supply that results when the average tax rate changes in the prescribed way but the marginal rate is held fixed at its initial value. (This is analogous to the shift in the budget constraint from Y_1N_1 to *bb* in fig. 3, which raises the average but not the marginal rate.) Subtracting the estimated labor supply under the demogrant policy from this yields the compensated change in labor supply. Then this change in labor supply is valued at the market wage rate to yield the compensated change in market earnings (*DC* in the diagram). From this figure the household's valuation of the compensated labor supply change must be subtracted. We estimated this by using a net wage rate halfway between the initial net marginal wage rate and the final one. For example, if the marginal tax rate rises from 40 to 41 percent, the household's value for the compensated labor supply change is 59.5 percent of its market valuation. Finally, the marginal welfare cost estimates are combined with each household's change in net tax revenues to determine the total loss or gain to the household.

Table 8 gives the results of these calculations, once again in the form of the average gain or loss per household in each quintile. As expected, in comparison to table 6 the gains to quintiles that benefit are larger while the losses to quintiles that are harmed are smaller. In the benchmark case, the income-equivalent loss to the top three quintiles is $3.49 per dollar of income-equivalent gain to the lowest quintiles, substantially less than the $9.51 trade-off rate when disposable money incomes alone are considered.

We believe these estimates suggest that the cost of moving further toward income equality by using tax and transfer programs is much higher than generally recognized. As evidence on this point, consider the words of the late Arthur Okun, who in *Equality and Efficiency: The Big Tradeoff* openly stated his views regarding the trade-off he would find acceptable. Okun likened moves to equalize incomes as transfers of money using a "leaky bucket": due to the adverse incentive effects the gain to recipients is less than the costs to taxpayers. Okun states (1975, p. 94) that he would accept a leakage of 60 percent for redistribution from the top 5 percent of households to the lowest 20 percent, but a leakage of only 15 percent for a redistribution from households with incomes 30 percent above the mean to 30 percent below

In view of Okun's expressed opinion that government sh
pand its redistributive activities, we are safe in concluding t
that actual marginal leakages fell short of these values. Ou
mates, however, suggest that actual marginal leakages are la
Okun would have accepted. Note that Okun's concept o
corresponds to (one minus) the reciprocal of our trade-off
example, in case *A* the ratio of losses to the top three quintiles to
gain for the bottom two quintiles is 3.49 and means that each dollar of cost to upper-income households only provides a benefit of 29 cents to low-income households, a leakage of 71 percent. Moreover, that should be compared to an average (perhaps weighted in some way) of Okun's 60 percent and 15 percent values since we consider the trade-off across all income classes and not just from the top to the bottom We note that the "leakages" implied by our estimates all exceed 70 percent, except for the first measure of the trade-off for cases *D* and *E*. Even in those cases, however, the "leakage" is about 56 percent, well in excess of the unweighted average value of 37.5 percent that

TABLE 8

EFFECTS OF 1 PERCENTAGE POINT INCREASE IN TAX RATE

	PARAMETERS		ELASTICITIES		INCOME-EQUIVALENT GAINS AND LOSSES					TRADE-OFF RATES	
CASE	α	β			Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5	Losses/ Gains	Losses to Top 4/ Gains to Bottom
A	2	3.25	.20	.31	$47.30	32.60	−11.49	−71.59	−196.21	3.49	5.22
B	.1	2.5	41	47	35.14	14.35	−32.45	−92.77	−213.39	6.84	9.23
C	3	2.5	31	47	38.55	18.07	−30.32	−93.45	−222.00	6.11	8.51
D	1	4.0	15	21	52.83	41.96	1.05	−56.50	−177.02	2.43	3.61
E	.3	4.0	.05	21	56.12	45.85	3.46	−56.80	−185.07	2.29	3.43

Okun found acceptable. In other words, given his expressed values, and assuming our estimates are reliable, Okun should have favored a reduced role for government in redistributing income.

Or consider Arnold Harberger's view: "While the most eminent minds in the economics profession would be hard put to find ways of effectuating quantitatively important transfers at efficiency costs of less than 1 percent of the amounts involved, most beginning graduate students could in one afternoon invent a hundred ways of bringing about major transfers at an efficiency cost of less than 20 percent of the amounts involved" (1978, p. S115). It seems almost certain that marginal welfare costs are much larger than Harberger suggests in this passage.

VI. Conclusion

Our major conclusion is that the marginal cost of less income inequality is surprisingly high even when labor supply elasticities are relatively low. This finding does not imply that we have necessarily gone beyond the "optimal" degree of income equality; it is just an attempt to identify the real trade-off involved. We do, however, hope this research will stimulate further attempts to investigate the relevant objective trade-offs that must be made to evaluate the government's redistributive role.

Given the narrow focus of the present paper on the labor supply effects of changes in tax-transfer policies, the overall trade-offs for actual policies may be greater or smaller. In addition to affecting labor supply, most actual redistributive policies involve other costs that we have ignored: administrative and compliance costs; effects on the composition of income and expenditures due to narrow measure of taxable income (the "loophole" problem); effects on family composition; effects on human capital accumulation; and c[illegible] consumption patterns due to the use of in-kind transfers. M[illegible] an important assumption in our model is that the marginal re[illegible]tion has no effect on saving. If, as seems likely, saving is [illegible] affected, then the marginal cost of reducing income inequality [illegible] substantially greater than our estimates imply (Mendelsohn 198[illegible]). Because of these omissions from the model, our presumption is that actual policies are more distorting than our policy that affects only the quantity of labor supplied; if this is so, our estimates would be lower bounds to the trade-off implied by actual policies.

On the other hand, it may be that a totally different type of policy could redistribute income at a lower cost. For example, wage rate subsidies or earnings subsidies (like the earned income tax credit) may be more efficient at the relevant margin than the type of negative income tax examined. In addition, structural changes in the present tax-transfer system might be capable of reducing the marginal trade-off.[18]

Finally, our estimates show that the magnitude of compensated supply elasticities has a major influence on the trade-off between equality and efficiency. Even for the relatively narrow range of compensated elasticities we considered, from 0.2 to 0.45, the trade-off showed wide variation. The importance of further empirical investigation of the determinants of labor supply in an effort to narrow the range of uncertainty is clearly underscored by this finding.

References

Aaron, Henry J., and Boskin, Michael J., eds. *The Economics of Taxation.* Washington: Brookings Inst., 1980.

Aaron, Henry J., and Pechman, Joseph A., eds. *How Taxes Affect Economic Behavior.* Washington: Brookings Inst., 1981.

Barro, Robert J., and Sahasakul, Chaipat. "Measuring the Average Marginal Tax Rate from the Individual Income Tax." Working Paper no. 1060. Cambridge, Mass.: N.B.E.R., 1983.

Baumol, William J., and Fischer, Dietrich. "The Output Distribution Frontier: Alternatives to Income Taxes and Transfers for Strong Equality Goals." *A.E.R.* 69 (September 1979): 514–25.

Bishop, John. "Latest SIME/DIME Results." Mimeographed. Madison: Univ. Wisconsin, 1980.

Browning, Edgar K. "The Marginal Welfare Cost of Income Redistribution." *Southern Econ. J.* 45 (July 1978): 1–17.

Browning, Edgar K., and Johnson, William R. *The Distribution of the Tax Burden.* Washington: American Enterprise Inst., 1979.

Council of Economic Advisers. *Economic Report of the President.* Washington: Government Printing Office, 1982.

Danziger, Sheldon; Haveman, Robert; and Plotnick, Robert. "How Income Transfer Programs Affect Work, Savings, and the Income Distribution: A Critical Review." *J. Econ. Lit.* 19 (September 1981): 975–1028.

Fullerton, Don. "On the Possibility of an Inverse Relationship between Tax Rates and Government Revenues." *J. Public Econ.* 19 (October 1982): 3–22.

Harberger, Arnold C. "On the Use of Distributional Weights in Social Cost-Benefit Analysis." *J.P.E.* 86, no. 2, pt. 2 (April 1978): S87–S120.

Hausman, Jerry. "Labor Supply." In *How Taxes Affect Economic Behavior,* edited by Henry J. Aaron and Joseph A. Pechman. Washington: Brookings Inst., 1981.

Heckman, James J.; Killingsworth, Mark; and MaCurdy, Thomas E. "Empirical Evidence on Static Labour Supply Models: A Survey of Recent Developments." In *Economics of the Labour Market,* edited by Zmira Hornstein, Joseph Grice, and Alfred Webb. London: H.M. Stationery Office, 1981.

---

[18] For example, reducing reliance on government for weakly redistributive policies such as social security would allow the trade-off to be improved. Our simulations have held constant all existing programs.

Keeley, Michael. *Labor Supply and Public Policy: A Critical Review.* New York: Academic Press, 1981.

Killingsworth, Mark. "A Survey of Labor Supply Models: Theoretical Analysis and First-Generation Empirical Results." In *Research in Labor Economics*, vol. 4, edited by Ronald G. Ehrenberg. Greenwich, Conn.: JAI, 1981.

Lampman, Robert. "Labor Supply and Social Welfare Benefits in the United States." In *Concept and Data Needs*, Appendix Vol. 1, by the National Commission on Employment and Unemployment Statistics. Washington: Government Printing Office, 1981.

MaCurdy, Thomas E. "An Intertemporal Analysis of Taxation and Work Disincentives." Working Paper no. E-80-4. Stanford, Calif.: Hoover Inst., Stanford Univ., 1980.

Mendelsohn, Robert. "The Choice of Discount Rates for Public Projects: Reply." *A.E.R.* 73 (June 1983): 499–500.

Mirrlees, James A. "An Exploration in the Theory of Optimum Income Taxation." *Rev. Econ. Studies* 38 (April 1971): 175–208.

Moffitt, Robert, and Kehrer, Kenneth. "The Effect of Tax and Transfer Programs on Labor Supply: Evidence from the Income Maintenance Experiments." In *Research in Labor Economics*, vol. 4, edited by Ronald G. Ehrenberg. Greenwich, Conn.: JAI, 1981.

Okun, Arthur M. *Equality and Efficiency: The Big Tradeoff.* Washington: Brookings Inst., 1975.

Rawls, John. *A Theory of Justice.* Cambridge, Mass.: Harvard Univ. Press, 1971.

Sadka, Efraim. "On Income Distribution, Incentive Effects and Optimal Income Taxation." *Rev. Econ. Studies* 43 (June 1976): 261–67.

Sheshinski, Eytan. "The Optimal Linear Income-Tax." *Rev. Econ. Studies* 39 (July 1972): 297–302.

Stern, Nicholas H. "On the Specification of Models of Optimum Income Taxation." *J. Public Econ.* 6 (July/August 1976): 123–62.

U.S. Congress, Congressional Budget Office. *Poverty Status of Families under Alternative Definitions of Income.* Background Paper no. 17. Washington: Government Printing Office, 1977.

The Effects on Output of Money Growth and Interest Rate Volatility in the United States

Paul Evans

Stanford University

In October 1979, the Federal Reserve began a policy of disinflation and also stopped stabilizing interest rates. Subsequently, interest rates became much more volatile. Using a methodology advanced by Barro, this paper estimates what this change in policy did to output in the United States. The paper finds that the stagnation of the U.S. economy since 1979 stemmed almost as much from increased interest rate volatility as from reduced money growth. The paper also assesses the extent to which increased volatility of money growth has contributed to stagnation. It finds no evidence of any appreciable contribution. The paper concludes that stabilizing interest rates is probably sensible monetary policy.

I. Introduction

During the 1970s, the central banks of most industrial countries became preoccupied with controlling inflation. Since long-term control over inflation requires long-term control over the money supply, many of these central banks have adopted some form of monetary targeting. Central banks, however, have generally regarded the targets they have set for money growth to be long-term targets. They have rarely responded much to short-term fluctuations in money growth. To have done so, in their view, would have sharply raised interest rate volatility with no compensating gain in their ability to wage war against inflation.[1]

I wish to thank James Heckman, Ronald McKinnon, Allan Meltzer, Tom Mayer, Robert Solow, Erich Streissler, Jose Vinals, and an anonymous referee for helpful comments.

[1] The articles in Meek (1983) detail the views of several central banks.

[*Journal of Political Economy*, 1984, vol. 92, no. 2]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0010$01.50

Monetarists have attacked these fluctuations in money growth, which they believe raise the variability of economic activity, contribute to uncertainty, and undermine the credibility of central banks. Because the Federal Reserve was somewhat persuaded by its monetarist critics and since it was losing its war against inflation, it embarked in October 1979 on a determined policy of disinflation.[2] Under the new policy, the Federal Reserve would actually try to hit its monetary targets in the short run. It would generally not intervene in securities markets to stabilize interest rates, thus allowing private market forces to determine them.

This change in policy has not satisfied monetarists. For example, Milton Friedman, in his *Newsweek* column of February 15, 1982, condemned the Federal Reserve for permitting "yo-yo" movements in money growth. Even though he felt that monetary policy had on average been about right since October 1979, these short-term fluctuations had produced great economic instability. Moreover, the public had been made uncertain about whether the policy of disinflation would last. Friedman continued his attack in *Newsweek* on August 23, 1982. In his column, he blamed the Federal Reserve for "monetary gyrations" that had "aborted" the recovery of the U.S. economy.

Meltzer (1982) launched one of the most vehement monetarist attacks to date. He called the Fed's policy of disinflation a "failed experiment" because it had produced an increased short-run variance of money growth, which, according to Meltzer, had resulted in uncertainty, disrupted financial markets, kept real interest rates high, and contributed in large measure to the recession.

Monetarists have not been the only economists to condemn Reserve policy since October 1979. At that time, the Federal not only began to pursue disinflation actively but also allowed rates to become much more volatile.[3] Many nonmonetarists this increased interest rate volatility has disrupted financial greatly increased the risk that economic agents operating markets must bear, and thus raised the risk premia in interest rates (see, e.g., B. Friedman 1982). Consequently, the new policy may have depressed the economy unnecessarily.

In this paper, I assess these two positions. I try to find out whether money growth and interest rate volatility affect the level of output Basically, I use a modification of the well-known model of Barro (1978, 1981). Barro's model is essentially monetarist in character, but my modification of it turns out to have nonmonetarist implications. I find no evidence that money growth volatility affects output at all. In contrast, interest rate volatility is an important determinant of output. In particular, the economy might not have stagnated since October 1979, had interest rates remained as stable as they were before 1979.

---

[2] The Fed came to believe that its short-run goal of interest rate stability conflicted with its goal of long-run monetary control. See the articles by Axilrod and Sternlight in Meek (1983)

[3] See my 1981 paper for evidence that some, but by no means all, of the increased volatility of interest rates has resulted from the change in operating procedures

The plan of this paper is as follows. Section II discusses Barro's model. Section III presents my measures of volatility for the interest rate and for unanticipated money growth. Section IV modifies Barro's model by including these volatility measures in his output equation. Section V extends the sample to 1981 and then uses the estimated model to determine why output has grown so slowly since 1978. Section VI expresses some reservations and makes an observation about the results. Finally, Section VII draws some conclusions.

II. Barro's Model for Output

The starting point for my empirical analysis is Barro's well-known model. Barro bases his model on the hypothesis that only unanticipated money growth affects real variables like output or the unemployment rate. This hypothesis is associated with the rational expectations school—Lucas, Sargent, Barro, etc.—but also is one with which many monetarists would agree.

To implement and test this hypothesis, one must devise some method for separating money growth into anticipated and unanticipated parts. Barro's method entails fitting an equation that explains money growth in terms of readily observable variables—past money growth, current fiscal policy (the public can tell when a war is going on), and the past unemployment rate. The fitted values from the regression are then the anticipated part of money growth, and the residuals are the unanticipated part. For example, in his 1981 book, Barro obtains[1]

$$\begin{aligned}\Delta \ln M_t = {} & \underset{(.015)}{.098} + \underset{(.09)}{.41}\Delta \ln M_{t-1} + \underset{(.07)}{.18}\Delta \ln M_{t-2} + \underset{(.009)}{.081}FEDV_t \\ & + \underset{(.005)}{.031} \ln\left[\frac{U_{t-1}}{(1 - U_{t-1})}\right] + DMR_t, \qquad (1) \\ & \text{S.E.} = .013,\ \text{D-W} = 1.8,\end{aligned}$$

using annual U.S. data over the sample period 1941–78. The figures in parentheses are standard errors. The variables in the equation are t, an index of time; M, the M1 money supply; $FEDV$, the proportional gap between real federal spending and its normal level;[5] U, the unemployment rate; and $\{DMR\}$, the residuals. Barro (1977) has rationalized the presence of each variable in the right-hand member of equation (1), so I need not do so here.

---

[1] This is eq (14) on p. 151 of Barro (1981). Note that the observations for 1941–46 received only 36 percent of the weight of the remaining observations.

Using the residual DMR as his measure of the unanticipated part of the money growth, Barro fits the output equation[6]

$$\begin{aligned}\ln Y_t = {} & \underset{(.03)}{2.88} + \underset{(.0004)}{.0829t} + \underset{(.19)}{1.01DMR_t} + \underset{(.22)}{.99DMR_{t-1}} \\ & + \underset{(.014)}{.081 \ln G_t} + EY_t,\end{aligned} \tag{2}$$

$$\text{S.E.} = .014, \text{ D-W} = 1.6,$$

to data for the sample period 1946–78, where G is real federal purchases of goods and services and $\{EY\}$ are the residuals from the equation. Since the coefficients on DMR and DMR lagged 1 year are positive and statistically significant at extremely low levels, Barro concludes that unanticipated money growth strongly affects output. Moreover, since adding current and lagged values of anticipated money growth ($\Delta \ln M - DMR$) to equation (2) does not raise its explanatory power by a statistically significant amount, Barro further concludes that only unanticipated money growth affects output as the theories of Lucas (1972), Barro (1976), and Sargent (1976) among others would suggest.

Germany and Srivastava (1979), Small (1979), Gordon (1980), and Mishkin (1982) have pointed out several problems with Barro's methodology and his implementation of it.[7] Nevertheless, for thre reasons, I shall use Barro's methodology in this paper. First, model is perhaps the most familiar simple model for outp macroeconomics literature. Second, Barro's model is one wi monetarists should be comfortable. Third, many nonmonetarists believe that Barro's model has at least some validity (see, e.g., Blinder 1980).

---

[5] More precisely, $FEDV_t = \ln (FED_t/FED_t^*)$, $\ln FED_t^* = 2\Sigma\, 8^i \ln FED$, equals nominal federal spending divided by the GNP deflator

[6] This is eq. (16) on p. 152 of Barro (1981) The sequence described in convenient for exposition, but in reality Barro estimated his equations simultaneous Simultaneous estimation is appropriate because, as Mishkin (1982) has shown sequential estimation reduces efficiency and biases the standard errors of the estimated coefficients

[7] In addition, Sims (1980*a*) has reported results from a vector autoregressive model, suggesting that unanticipated money growth has little explanatory power in the postwar period once one includes interest rates in the model King (1982), however, has shown that this result is not robust to changes in specification When the variables in the model are differenced or when time trends are included in each equation of the model, unanticipated money growth becomes important again Litterman and Weiss (1983) have found evidence that the transmission process between nominal interest rates and output does not pass through real interest rates. Later in the paper, I report evidence that Barro's definition of unanticipated money growth dominates nominal interest rates in explaining annual output movements.

While I wish to use Barro's methodology, I do not want to use equations (1) and (2) as they now stand, for several changes are in order. For one thing, money growth has apparently behaved much differently since 1947 from how it behaved before then. For evidence, see equations (3) and (4) below, which I fitted to quarterly data similar to those used by Barro and Rush (1980):

$$\begin{aligned}\Delta \ln \hat{M} = {} & \underset{(.0136)}{-.0006} - \underset{(.184)}{.055}\Delta \ln M_{t-1} + \underset{(.179)}{.372}\Delta \ln M_{t-2} \\ & - \underset{(.169)}{.236}\Delta \ln M_{t-3} - \underset{(.187)}{.407}\Delta \ln M_{t-4} + \underset{(.0112)}{.0512}FEDV_t \\ & + \underset{(.0182)}{.0035} \ln \left[\frac{U_{t-1}}{(1 - U_{t-1})}\right] - \underset{(.0206)}{.0171} \ln \left[\frac{U_{t-2}}{(1 - U_{t-2})}\right] \\ & + \underset{(.0242)}{.0645} \ln \left[\frac{U_{t-3}}{(1 - U_{t-3})}\right], \\ & R^2 = .797,\ \text{S.E.} = .01322,\ \text{D-W} = 2.50,\end{aligned} \tag{3}$$

and

$$\begin{aligned}\Delta \ln \hat{M}_t = {} & \underset{(.00105)}{.00080} + \underset{(.092)}{.698}\Delta \ln M_{t-1} - \underset{(.108)}{.081}\Delta \ln M_{t-2} \\ & + \underset{(.096)}{.184}\Delta \ln M_{t-3} + \underset{(.080)}{.004}\Delta \ln M_{t-4} + \underset{(.00397)}{.00582}FEDV_t \\ & - \underset{(.0047)}{.0020} \ln \left[\frac{U_{t-1}}{(1 - U_{t-1})}\right] + \underset{(.0049)}{.0114} \ln \left[\frac{U_{t-2}}{(1 - U_{t-2})}\right] \\ & + \underset{(.0048)}{.0152} \ln \left[\frac{U_{t-3}}{(1 - U_{t-3})}\right], \\ & R^2 = .575,\ \text{S.E.} = .004873,\ \text{D-W} = 1.98.\end{aligned} \tag{4}$$

Equation (3), which I fitted over the sample period 1941II–1946IV, differs markedly from equation (4), which I fitted over the sample period 1947I–1978IV. For this reason, it makes no sense to fit the same equation to the 1941–46 observations as to the 1947–78 observations. I therefore fitted the following equations, which are the best analogues to equation (1) that I could find:

$$\underset{}{\Delta \ln M_t} = \underset{(.0279)}{.0767} + \underset{(.0266)}{.0962 FEDV_t} + DMR_t, \tag{5a}$$

$$1941\text{–}46,\ R^2 = .7656,\ \text{S.E.} = .03424,\ \text{D-W} = 3.32,^{8}$$

and

$$\begin{aligned}\Delta \ln M_t &= \underset{(.0223)}{.0525} + \underset{(.108)}{.693\Delta \ln M_{t-1}} + \underset{(.0163)}{.0600 FEDV_t} \\ &\quad + \underset{(.0114)}{.0332} \ln \left[\frac{U_{t-1}}{(1 - U_{t-1})}\right] - \underset{(.0064)}{.0166} \ln \left[\frac{U_{t-2}}{(1 - U_{t-2})}\right] + DMR_t\end{aligned} \tag{5b}$$

$$1947\text{–}78,\ R^2 = .6896,\ \text{S.E.} = .01379,\ \text{D-W} = 2.00.^{9}$$

These equations differ appreciably from equation (1). The differences stem primarily from splitting the sample, although the revisions in nominal federal spending, the GNP deflator, and the money supply do make some difference. In addition, I have used the civilian unemployment rate for U in lieu of Barro's measure, the unemployment rate of civilians plus the military. Since the two differ only slightly after 1946, this change makes little difference.

Another reason why I do not accept equations (1) and (2) is that the residuals in equation (2) are too serially correlated (D-W = 1.6). Moreover, the serial correlation becomes much worse when I replace Barro's *DMR* with the residuals from equations (5a) and (5b). The serial correlation markedly improves, however, when I include ln Y lagged 1 year in the right-hand member.

I fitted the analogue of equation (2), obtaining[10]

$$\begin{aligned}\ln Y_t &= \underset{(1.03)}{3.54} + \underset{(.0067)}{.0212t} + \underset{(.194)}{.365 \ln Y_{t-1}} + \underset{(.320)}{2.152 DMR} \\ &\quad + \underset{(.392)}{1.142 DMR_{t-1}} + \underset{(.0328)}{.0356 \ln G_t} + EY_t.\end{aligned}$$

$$1947\text{–}78,\ R^2 = .9952,\ \text{S.E.} = .02512,\ \text{D-W} = 2.00$$

Equation (6) differs from equation (2) in two important ways. First, equation (6) includes a lagged dependent variable but equation (2) does not. Output therefore undergoes persistent movements that

---

[8] This Durbin-Watson statistic does not indicate statistically significant serial correlation because there are only 4 degrees of freedom

[9] Chow tests revealed that the coefficients of eq. (5b) did not vary by a statistically significant amount over the sample period

[10] Chow tests revealed that the coefficients of eq. (6) did not vary by a statistically significant amount over the sample period.

stem from causes other than unanticipated money growth. Second, equation (6) implies a much larger contemporaneous effect for unanticipated money growth and a somewhat smaller effect for real federal purchases than does equation (2). Moreover, real federal purchases no longer have an effect that is statistically significant at conventional levels.

In fitting equations (5a) and (6), I used the full information maximum likelihood program in TSP. As pointed out in note 6, joint estimation is necessary to obtain efficient estimates of the coefficients and consistent estimates of the standard errors. I have also used this program to estimate all equations reported below.

Barro never intended his model to answer the questions central to this paper. For that reason, he did not include in his output equation any measures of how volatile money growth and interest rates are. In the next section, I develop such measures. I then modify Barro's output equation by including them in it.

III. Measures of Volatility

If equation (5b) should be specified completely correctly, one cannot measure how changes in the volatility of unanticipated money growth affect output. The reason is that this volatility would then be a constant—say, the standard error of equation (5b). In reality, the error variance of equation (5b) is probably not entirely constant. My parameter estimates are therefore consistent, but not efficient. If one could parameterize how the error variance changed over time, one can restore efficiency and also test whether the volatility of unanticipated annual growth affects output. Unfortunately, one probably cannot. What I have done instead is to ignore the inefficiency of equation (5b), measuring the volatility of unanticipated money growth as $VM_t = [1/6 \sum_{i=0}^{5} DMR_{t-i}^2]^{1/2}$. In other words, I measured it as the 6-year moving standard deviation of unanticipated money growth.[11] In using VM as my measure of money growth volatility, I am implicitly assuming that, because anticipated money growth has no output effects, any volatility of money growth stemming from anticipated changes in money growth also has no output effects. Strictly speaking, VM should be termed the unpredictability of money growth rather than its volatility.

In addition to VM, I shall also consider $VR_t = [1/12 \sum_{i=1}^{12} r_{it}^2 - \bar{r}_t^2]^{1/2}$, where VR is the standard deviation of r_{it}, the monthly change in the AAA bond rate in month i of year t, and $\bar{r}_t = 1/12 \sum_{i=1}^{12} r_{it}$. In what

---

[11] I obtained essentially the same results using 3-, 4-, and 5-year standard deviations in lieu of VM

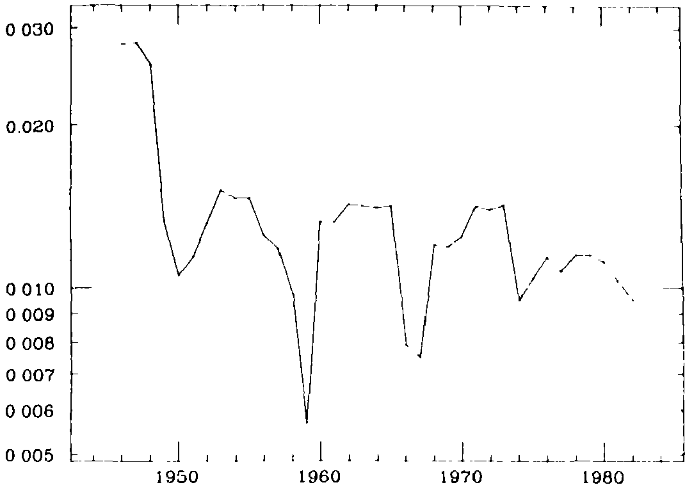


FIG. 1.—The volatility of unanticipated money growth

follows, I shall call *VR* the volatility of the interest rate.[12] Note that *VR* includes all interest rate volatility, whether it be predictable or not. Sargent (1976), however, has shown that, to a first approximation, changes in long-term bond rates are completely unpredictable. Therefore, *VR* also closely approximates a measure of the unpredictability of the interest rate.

Figures 1 and 2 plot these measures of volatility. The volatility of unanticipated money growth averaged 1.28 percent over the period 1947–81. It fell from its peak of 2.81 percent in 1947 to 1.32 perce
in 1949 and then, for the most part, stayed near that level for
of the postwar period. In particular, it has fallen, not risen, s
Federal Reserve adopted its new monetary policy in Octobe
contrary to the views expressed by Friedman and Meltzer
years that stand out since 1949 are 1959, 1966, and 1967 w
was unusually low.

Between 1947, when the Federal Reserve was still supporting bon prices, and 1958, when it no longer was, interest rate volatility followed a jagged, upward-sloping path. In the next 6 years, this upward trend was reversed. Interest rate volatility was as low in 1963–65 as it has ever been in the postwar period. *VR* then soared until 1969

---

[12] Strictly speaking, *VR* is long-term interest rate volatility. In the regressions reported in the next section, I tried including a measure of short-term interest rate volatility as well as *VR*. I found no statistically significant explanatory power for this variable.

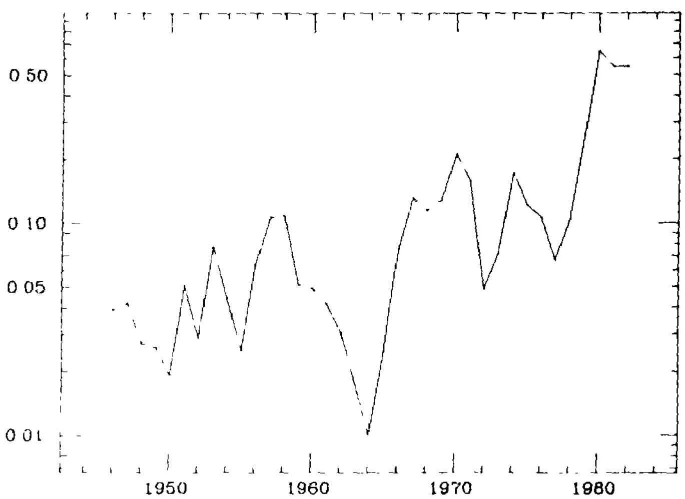


FIG. 2.—The volatility of the interest rate

During the next 10 years, VR averaged about .10, a value comparable to its peak value in the 1950s. It oscillated about that average with considerable amplitude, however. Finally, following the change in monetary policy in October 1979, VR shot upward. In 1980 and 1981, it averaged .60, which is much larger than .08, its average between 1947 and 1978.

IV. Modification of Barro's Output Equation

I first estimated an output equation that included ln VM_t, ln VM_{t-1}, ln VR_t, ln VR_{t-1}, and ln VR_{t-2} as well as t, ln Y_{t-1}, DMR_t, DMR_{t-1}, and ln G_t.[13] I have reported the estimated coefficients in the first row of table 1. Since ln VR_t is not statistically significant at any reasonable level and since it is the least significant variable, I estimated another output equation that excludes this variable. The second row of table 1 reports the coefficients obtained. Dropping ln VM_t, the least significant variable in the equation, and reestimating yielded the coefficients reported in row 3. The coefficients in the last row come from dropping ln VM_{t-1}, the variable with the least significant coefficient in row 3. All of the coefficients reported in the last row are either highly significant or else suggestive, and the error term appears to be serially uncor-

---

[13] The logarithm is the natural transform to apply to any variable ranging over the open interval (0, ∞).

TABLE 1

VARIOUS MODIFICATIONS OF BARRO'S OUTPUT EQUATION

Constant	t	$\ln Y_{t-1}$	DMR_t	DMR_{t-1}	$\ln G_t$	$\ln VM_t$	$\ln VM_{t-1}$	$\ln VR_t$	$\ln VR_{t-1}$	$\ln VR_{t-2}$	R^2	S.E.	D-W
3.47	.0217	.358	2 352	956	0365	- 0024	- 0139	.0010	- 0191	0084	.9963	02433	1.96
(1.36)	(.0093)	(255)	(309)	(252)	(0312)	(0120)	(.0153)	(.0099)	(0138)	(0084)			
3.40	.0213	.370	2.343	956	0360	- 0028	- 0144		- 0186	0085	.9963	.02374	1.96
(.90)	(0062)	(174)	(283)	(259)	(.0282)	(0092)	(0150)		(.0067)	(0071)			
3.40	.0213	372	2 346	956	0356		- 0153		- 0189	.0084	.9963	.02333	1.96
(.88)	(0061)	(.169)	(.280)	(251)	(0280)		(0142)		(0065)	(.0069)			
3.38	.0212	.385	2 370	998	0355				- 0164	.0049	.9962	.02356	2.01
(90)	(.0061)	(171)	(304)	(.267)	(0282)				(0050)	(.0044)			

NOTE —Standard errors appear in parentheses below each estimated coefficient The output equations were estimated over the sample period 1947–78, using full information maximum likelihood related.[14] Finally, Chow tests revealed that the coefficients did not shift significantly over the sample period.

The coefficients on VR seem to imply that a change in interest rate volatility can greatly affect output. For example, if VR should rise by 100 percent some year (approximately what happened in 1980 and 1981), output is reduced by 1.64 percent the next year, by 1.64 × 1.385 − 0.49, or 1.78, percent the year after next, and by (1.64 − 0.49)/(1 − .385), or 1.87, percent ultimately. This interpretation, however, is probably mistaken. To see why, consider the equations below:

$$
\begin{aligned}
\ln Y_t = {} & \underset{(1.12)}{3.04} + \underset{(.0074)}{.0191t} + \underset{(.221)}{.445} \ln Y_{t-1} + \underset{(.228)}{2.395} DMR_t \\
& + \underset{(.345)}{.903} DMR_{t-1} + \underset{(.0386)}{.0327} \ln G_t - \underset{(.0072)}{.0167} \ln VR_{t-1} \\
& + \underset{(.0074)}{.0083} \ln VR_{t-2} - \underset{(.0063)}{.0043} \ln VR_{t-3} + EY_t, \\
& 1947\text{–}78,\ R^2 = .9960,\ \text{S.E.} = .02443,\ \text{D-W} = 2.05,
\end{aligned}
\tag{7}
$$

and

$$
\begin{aligned}
\ln VR_t = {} & -\underset{(1.08)}{2.37} + \underset{(.0193)}{.0300t} + \underset{(.212)}{.738} \ln VR_{t-1} - \underset{(.267)}{.321} \ln VR_{t-2} + EVR_t, \\
& 1947\text{–}78,\ R^2 = .5412,\ \text{S.E.} = .5343,\ \text{D-W} = 2.10.
\end{aligned}
\tag{8}
$$

Notice that .0083/.0167 (= .497) and −.0043/.0167 (= −.257) do not differ much from the coefficients of $\ln VR_{t-1}$ and $\ln VR_{t-2}$ in equation (8). It is therefore plausible that $\ln VR_{t-1}$, $\ln VR_{t-2}$, and $\ln VR_{t-3}$ appear in equation (7) because output depends on EVR_{t-1}, last year's unanticipated component of interest rate volatility. Indeed, one cannot reject this hypothesis at any reasonable significance level. Imposing the restriction that only the unanticipated component matters yields

$$
\begin{aligned}
\ln Y_t = {} & \underset{(1.11)}{2.81} + \underset{(.0073)}{.0168t} + \underset{(.220)}{.494} \ln Y_{t-1} + \underset{(.332)}{2.505} DMR_t \\
& + \underset{(.349)}{.854} DMR_{t-1} + \underset{(.0380)}{.0325} \ln G_t - \underset{(.0079)}{.0173} EVR_{t-1} + EY_t, \\
& 1947\text{–}78,\ R^2 = .9956,\ \text{S.E.} = .02448,\ \text{D-W} = 2.05.
\end{aligned}
$$

---

[14] In calling the last-row coefficients statistically significant, I am ignoring the tendency of pretesting to raise the apparent significance of coefficients above their true significance. None of the first 10-sample autocorrelations of the residuals is statistically significant at the .05 level.

Because ln G is not statistically significant in this equation at any reasonable significance level, I reestimated the equation, excluding real federal purchases. The result is

$$\begin{aligned}\ln Y_t = {} & \underset{(.83)}{2.41} + \underset{(.0050)}{.0136t} + \underset{(.142)}{.597 \ln Y_{t-1}} + \underset{(.438)}{2.544 DMR_t} \\ & + \underset{(.307)}{.923 DMR_{t-1}} - \underset{(.0077)}{.0178 EVR_{t-1}} + EY_t,\end{aligned} \tag{9}$$

$$1947\text{–}78,\ R^2 = .9951,\ \text{S.E.} = .02523,\ \text{D-W} = 2.03.$$

Every coefficient in this equation is highly significant, and the error term exhibits no obvious serial correlation. For future reference, I also report the equations for money growth and for interest rate volatility that I estimated jointly with equation (9):

$$\begin{aligned}\Delta \ln M_t = {} & \underset{(.0255)}{.0552} + \underset{(.083)}{.780 \Delta \ln M_{t-1}} + \underset{(.0151)}{.0392 FEDV_t} \\ & + \underset{(.0134)}{.0300} \ln \left[\frac{U_{t-1}}{(1 - U_{t-1})}\right] - \underset{(.0071)}{.0124} \ln \left[\frac{U_{t-2}}{(1 - U_{t-2})}\right] + DMR_t,\end{aligned} \tag{10}$$

$$1947\text{–}78,\ R^2 = .6622,\ \text{S.E.} = .01438,\ \text{D-W} = 2.11,$$

and

$$\ln VR_t = \underset{(.74)}{-2.64} + \underset{(.0155)}{.0333t} + \underset{(.155)}{.690 \ln VR_{t-1}} - \underset{(.177)}{.338 \ln VR_{t-2}} + EVR_t, \tag{11}$$

$$1947\text{–}78,\ R^2 = .5386,\ \text{S.E.} = .5358,\ \text{D-W} = 2.00.$$

Equation (9) has a substantially different interpretation from th one that I tentatively gave the last row of table 1. Output res great deal to *unanticipated* interest rate volatility. For exampl anticipated increase of 100 percent in interest rate volatility next year's output by 1.78 percent. In contrast, *anticipated* ch interest rate volatility do not affect output at all.[15]

Sims (1980*a*) has argued that interest rates are primarily w fects output, not unanticipated money growth. In a conversation Robert Solow has suggested that VR in table 1 and equation (7) and EVR in equation (9) are proxying for the level of interest rates. To test these hypotheses, I have included in regression equations of the form (9) current and past values of the annual averages of the AAA bond rate. Table 2 reports the test statistics for including these additional regressors. It also reports some critical values for the test statistics.

---

[15] The same result holds if eq. (9) is fitted in the first differences of its variables. The anticipated part of ln VR has no statistically significant explanatory power

TABLE 2

STATISTICS FOR TESTING WHETHER INTEREST RATES HAVE DIRECT EFFECTS ON OUTPUT

LAGS INCLUDED	TEST STATISTICS*	*df*	CRITICAL VALUES 10%	CRITICAL VALUES 5%
0, 1, 2, 3	6.13	4	7.78	9.49
0, 1, 2	4.66	3	6.25	7.81
0, 1	2.65	2	4.61	5.99
0	2.58	1	2.71	3.84
1, 2, 3	5.34	3	6.25	7.81
1, 2	3.68	2	4.61	5.99
1	1.64	1	2.71	3.84

* In large samples, the test statistics are approximately distributed as χ^2 with the indicated degrees of freedom.

Clearly, the AAA bond rate adds no statistically significant explanatory power. Moreover, the estimated coefficients on DMR_t, DMR_{t-1}, and EVR_{t-1} and their standard errors do not change much. Therefore, the annual data analyzed here validate neither Sims's nor Solow's hypothesis.

Lucas (1973) has argued that output responds more to unanticipated money growth, the less variable nominal shocks are and the more variable real shocks are. This hypothesis implies that the coefficients on DMR_t and DMR_{t-1} in equation (9) should be decreasing functions of money growth volatility and increasing functions of interest rate volatility.[16] I have therefore fitted the analogue of equation (9) with the coefficients on DMR_t and DMR_{t-1} replaced by variable coefficients of the form $a_0 + a_1 \ln VM_t + a_2 \ln VR_t$, where $a_1 < 0$, $a_2 > 0$, and $b_0 + b_1 \ln VM_t + b_2 \ln VR_t$, where $b_1 < 0$, $b_2 > 0$. I found that one cannot reject the joint null hypothesis $a_1 = a_2 = b_1 = b_2 = 0$ at any reasonable significance level. One also cannot reject the null hypotheses $a_1 = 0$, $a_2 = 0$, $b_1 = 0$, $b_2 = 0$, or any combination thereof. I have therefore not altered the specification (9).

Money growth volatility does not appear in equation (9), but unanticipated interest rate volatility does.[17] For that reason, it seems plausible that the Federal Reserve should pay considerable attention to interest rate volatility even if doing so were to raise money growth volatility somewhat.

---

[16] I assume here that VR is more related to the variance of real shocks than to the variance of nominal shocks.

[17] Since Friedman and Meltzer have emphasized the volatility of money growth over a shorter term than 1 year, I also tried measures of money growth volatilities for terms of 1, 2, 3, 4, and 6 months. None of these measures had statistically significant explanatory power at any conventional level

V. Effects on Output of the October 1979 Change in Monetary Policy

Thus far, I have used data only through 1978. I limited the samples to these data because Barro fitted his model to them and because the change in monetary policy in October 1979 may have altered the parameters of the model as Lucas, Sargent, Barro, and others might argue. I now investigate whether the model did in fact change after 1978. First, consider the following analogues of equations (9)–(11), fitted over the sample period 1947–81:[18]

$$\begin{aligned}\ln Y_t = {} & \underset{(.78)}{2.36} + \underset{(.0047)}{.0132t} + \underset{(.133)}{.606} \ln Y_{t-1} + \underset{(.424)}{2.754}DMR_t \\ & + \underset{(.282)}{.901}DMR_{t-1} - \underset{(.0072)}{.0188}EVR_{t-1} + EY_t,\end{aligned} \qquad (12)$$

$$1947\text{–}81,\ R^2 = .9953,\ \text{S.E.} = .02660,\ \text{D-W} = 2.07;$$

$$\begin{aligned}\Delta \ln M_t = {} & \underset{(.0235)}{.0558} + \underset{(.076)}{.782} \ln M_{t-1} + \underset{(.0144)}{.0363}FEDV_t \\ & + \underset{(.0125)}{.0286} \ln\left[\frac{U_{t-1}}{(1 - U_{t-1})}\right] - \underset{(.0068)}{.0109} \ln\left[\frac{U_{t-2}}{(1 - U_{t-2})}\right] + DMR_t,\end{aligned} \qquad (13)$$

$$1947\text{–}81,\ R^2 = .6957,\ \text{S.E.} = .01391,\ \text{D-W} = 2.13;$$

and

$$\ln VR_t = \underset{(.66)}{-2.58} + \underset{(.0139)}{.0326t} + \underset{(.144)}{.691} \ln VR_{t-1} - \underset{(.161)}{.324} \ln VR_t$$

$$1947\text{–}78,\ R^2 = .5397,\ \text{S.E.} = .5087,\ \text{D-W} = 2.01$$

These equations do not differ appreciably from equations thereby suggesting that the parameters changed little, if at residuals of equations (12) and (13) for 1979, 1980, and 1981 much smaller than their standard errors. Consequently, equations (12) and (13) fit about as well in the period 1979–81 as in the period 1947–78. A final piece of evidence is that the dynamic forecasts from equations (12) and (13) stayed fairly close to the actual money growth rates and output levels for 1979, 1980, and 1981.

---

[18] Later in this section I shall present strong evidence that eq. (11) changed in 1979. In estimating eqq. (12) and (13) below, I assumed that the public has continued to anticipate interest rate volatility as they did in the sample period 1947–78. In other words, I used only pre-1979 data in estimating eq. (14); the 1-year-ahead prediction errors of eq. (14) served as the *EVR*'s for 1979–82.

	DMR	*EY*
1979	.00225	−.00461
1980	−.00926	.00298
1981	−.00146	.00189
S.E.	.01391	.02660

Since the model appears not to have changed in 1979, one can use equation (12) to analyze why the U.S. economy has performed so poorly since 1979 and to forecast what happened in 1982. According to equation (12), output tends to rise 1.32/(1 − .606), or 3.35, percent a year if all money growth and interest rate volatility have always been anticipated and if all error terms $\{EY\}$ have always been zero. Because these conditions have not been met, output has often departed from this trend line. The table below presents the percentage fall in output after 1978. It decomposes the departure into four components: that due to nonzero *DMR*'s in 1979, 1980, and 1981; that due to nonzero *EVR*'s in 1979 and 1980; that due to nonzero residuals in 1979, 1980, and 1981; and that due to the initial conditions in 1979. The table makes it clear that output has stagnated since 1979 as much because interest rates have been unusually volatile as because the Federal Reserve's policy of disinflation has lowered money growth below what the public anticipated. Moreover, since *EVR* was 0.48 in 1981 and 0.85 in 1982, output was reduced by 2.47 percent in 1982 and can be predicted to be reduced by 3.09 percent in 1983.

Year	*DMR*	*EVR*	*EY*	Initial Conditions	Total
1979	−.62	.00	.46	.71	.55
1980	1.97	1.14	.02	1.18	4.27
1981	2.43	2.59	−.20	.88	5.70

Equations (12) and (13) fit the 1979–81 observations as well as they fit the 1947–78 observations. In contrast, the interest rate volatilities for 1979–82 are much higher than one would have predicted using equation (14), which I fitted to 1947–78 observations. The table below reports the 1-year-ahead prediction errors for *VR*. If these prediction

Year	*EVR*
1979	.60
1980	1.01
1981	.48
1982	.85
Average	.74

errors were normally and independently distributed and had the same standard deviation as the 1947–78 residuals of equation (14) do, the odds against their averaging at least 0.74 would be more than 300:1.

Interest rates have become so much more volatile in large part because the Federal Reserve adopted new operating procedures.[19] Before October 1979, the Federal Reserve pegged the federal funds rate within narrow bands for appreciable periods of time. Shifts in the private demands for or supplies of money and credit produced accommodating changes in nonborrowed reserves and little change in market interest rates such as the AAA bond rate. Consequently, interest rate volatility was low. Since October 1979, the Federal Reserve has instead used nonborrowed reserves as its policy instrument, focusing its attention primarily on controlling money growth. Under this regime, shifts in supply or demand affect market interest rates appreciably because nonborrowed reserves no longer serve as a buffer. Consequently, interest rate volatility has become high.

If the new operating procedures have produced increased interest rate volatility that the public has not anticipated, adopting the new operating procedures was clearly a costly error. Eventually, however, one would expect that the public will come to anticipate higher interest rate volatility if the Federal Reserve continues to use its current operating procedures. These costs will then be one-time, sunk costs that need not then concern the Federal Reserve.[20]

VI. Some Reservations and an Observation

Equations (12)–(14) are most charitably regarded as three reduced
forms from some structural model. The results that I have [illegible]
in previous sections would be more credible had I derived th
a correct structural model of the U.S. economy rather th
these estimated reduced forms. From the structural model,
have learned which decisions of what agents depend on my
measures.[21] As it is, I have reduced forms whose structural
tation is unclear.

---

[19] Mascaro and Meltzer (1983) also reach this conclusion. They also argue that the new operating procedures have created short-term money growth volatility and that this money growth volatility together with interest rate volatility have raised real interest rates and depressed the economy.

[20] This discussion implicitly assumes that output adequately proxies for welfare. It also assumes that the variance of *EVR* does not depend on the operating procedures chosen by the Federal Reserve.

[21] Using a measure of interest rate volatility similar to mine, Starr (1982) may have answered part of this question. He shows that corporations shift toward short-term debt as his measure of interest rate volatility rises.

Of course, it takes a leap of faith to believe that one can identify a general-equilibrium structural model. Sims (1980*b*) has shown that the restrictions employed in identification are generally incredible. Reduced forms may therefore provide the only reliable information on how the economy behaves. Nevertheless, I should at some point try to fit a structural model and then use it to assess my results. Until I do, I shall have some reservations about the conclusions that I draw in the next section. The fitting of this model, however, must await another paper.

In a model with frictionless capital markets and no bond illusion, the monetary authorities cannot affect any allocative decision merely by stabilizing interest rates. In that kind of world, what matters for private allocative decisions, as Modigliani and Miller (1958) showed long ago, is various attributes of the earnings streams generated by actual capital assets. By speculating in some subset of the claims on these earnings streams, the Federal Reserve cannot affect anything because the resulting profits and losses of the central bank show up again in private earnings streams. An appropriate change in portfolio composition then permits every private agent to obtain exactly the same earnings stream it had without speculation.

Real-world capital markets are not frictionless, and the public may suffer from bond illusion, at least temporarily. Furthermore, information is not costless to obtain and to evaluate. For these reasons, one may reasonably expect that unpredictable changes in how the Federal Reserve speculates in securities markets will affect private allocative decisions. On the other hand, the Modigliani-Miller theorem may well apply to predictable changes, at least approximately. Equation (12) is consistent with this analysis. Output, which results from private allocative decisions, depends on the unanticipated component of interest rate volatility and not on the anticipated component.

IX. Conclusions

I have found evidence for the following conclusions: (1) Money growth volatility does not affect output. (2) Unanticipated interest rate volatility reduces output. (3) Output does not depend on anticipated interest rate volatility. (4) The new operating procedures adopted by the Federal Reserve in October 1979 led to an unanticipated increase in interest rate volatility, thereby reducing output by about 1 percent in 1980 and by about 2.5 percent in 1981 and 1982.

These conclusions contradict some tenets of monetarism.[22] They also clearly suggest that the new operating procedures of the Federal Reserve have been an unmitigated disaster. The obvious policy recommendation is that the Federal Reserve should begin again to use the federal funds rate as its policy instrument and to make interest rate stability one of its policy goals.[23]

---

[22] My findings, however, do not challenge the central tenets of monetarism. In a private communication, Allan Meltzer has agreed with this assessment and has expressed some sympathy for the approach of the paper.

References

Barro, Robert J. "Rational Expectations and the Role of Monetary Policy." *J. Monetary Econ.* 2 (January 1976): 1–32.

———. "Unanticipated Money Growth and Unemployment in the United States." *A.E.R.* 67 (March 1977): 101–15.

———. "Unanticipated Money, Output, and the Price Level in the United States." *J.P.E.* 86 (August 1978): 549–80.

———. *Money, Expectations, and the Business Cycle.* New York: Academic Press, 1981.

Barro, Robert J., and Rush, Mark. "Unanticipated Money and Economic Activity." In *Rational Expectations and Economic Policy,* edited by Stanley Fischer. Chicago: Univ. Chicago Press (for N.B.E.R.), 1980.

Blinder, Alan S. "Comment." In *Rational Expectations and Economic Policy,* edited by Stanley Fischer. Chicago: Univ. Chicago Press (for N.B.E.R.), 1980.

Evans, Paul. "Why Have Interest Rates Been So Volatile?" *Fed. Reserve Bank of San Francisco Econ. Rev.* (Summer 1981), pp. 7–20.

Friedman, Benjamin M. "Federal Reserve Policy, Interest Rate Volatility, and the U.S. Capital Raising Mechanism." *J. Money, Credit and Banking* 14, no. 4, pt. 2 (November 1982): 721–45.

Friedman, Milton. "The Yo-Yo Economy." *Newsweek* 96 (February 15, 1982), 72.

———. "An Aborted Recovery?" *Newsweek* 98 (August 23, 1982): 59.

Germany, J. David, and Srivastava, S. "Empirical Estimates of Unanticipated Policy: Issues in Stability and Identification." Unpublished manu M.I.T., 1979.

Gordon, Robert J. "Comment." In *Rational Expectations and Econom* edited by Stanley Fischer. Chicago: Univ. Chicago Press (for N 1980.

---

[23] In a letter, Tom Mayer has suggested that abandoning the federal funds policy instrument and the stabilization of interest rates as one of its policy go necessary for the policy of disinflation to be successful. His point is essentially that ev hough the funds rate is technically superior to nonborrowed reserves as a policy nstrument (Timsley et al. 1981) and even if the stabilization of interest rates is a desirable policy goal, the political environment makes the use of the funds rate and the pursuit of *stable* interest rates equivalent to the pursuit of *low* interest rates. Since nflation became chronic by 1979 because interest rates had been kept below their natural rates in the 1960s and 1970s, interest rates had to rise for disinflation to occur. Therefore, if Mayer is right, one should look at the output losses that I have measured s part of the price that had to be paid for disinflation to succeed. In this case, my policy conclusion would be wrong. Note that the new operating procedures only temporarily mpose an output loss since *EVR* must average zero. Because the benefits to a permanently lowered inflation rate last forever, their present value is likely to exceed the costs f this output loss.

King, Stephen R. "Interest Rates and the Transmission of Monetary Fluctuations to Output and Prices." Unpublished manuscript, Northwestern Univ., 1982.

Litterman, Robert B., and Weiss, Laurence. "Money, Real Interest Rates, and Output: A Reinterpretation of Postwar U.S. Data." Working Paper no. 1077, N.B.E.R., February 1983.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24.

———. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34.

Mascaro, Angelo, and Meltzer, Allan H. "Long- and Short-Term Interest Rates in a Risky World." *J. Monetary Econ.* 12 (November 1983): 485–518.

Meek, Paul. *Central Bank Views on Monetary Targeting.* New York: Fed. Reserve Bank of New York, 1983.

Meltzer, Allan H. "The Results of the Fed's Failed Experiment." *Wall Street J.* 108 (July 29, 1982): 24.

Mishkin, Frederic S. "Does Anticipated Aggregate Demand Policy Matter? Further Econometric Results." *A.E.R.* 72 (September 1982): 788–802.

Modigliani, Franco, and Miller, Merton H. "The Cost of Capital, Corporation Finance and the Theory of Investment." *A.E.R.* 48 (June 1958): 261–97.

Sargent, Thomas J. "A Classical Macroeconometric Model for the United States." *J.P.E.* 84 (April 1976): 207–38.

Small, David H. "Unanticipated Money Growth and Unemployment in the United States: Comment." *A.E.R.* 69 (December 1979): 996–1003.

Sims, Christopher A. "Comparison of Interwar and Postwar Business Cycles: Monetarism Reconsidered." *A.E.R. Papers and Proc.* 70 (May 1980): 250–57. (*a*)

———. "Macroeconomics and Reality." *Econometrica* 48 (January 1980): 1–48. (*b*)

Starr, Ross M. "Variation in the Maturity Structure of Debt and Behavior of the Monetary Aggregates: The Maturity Shift Hypothesis." Discussion Paper, Dept. Econ., Univ. California, San Diego, 1982.

Tinsley, Peter A.; von zur Muehlen, Peter; Trepeta, W.; and Fries, G. "Money Market Impacts of Alternative Operating Procedures." In *New Monetary Control Procedures.* Federal Reserve Staff Study. Vol. 2. Washington: Fed. Reserve Bd. of Governors, 1981.

Crime on the Court

Robert E. McCormick and Robert D. Tollison

Clemson University

This paper addresses the question, What happens to the arrest rate when the number of law enforcers increases? One implication of the analysis is that arrest statistics are a poor instrumental variable for judging the quality of law enforcement. Increasing the number of police can increase or decrease the number of arrests. An increased probability of arrest induces fewer criminal acts; hence the ambiguity. Because of this result, we apply the theory in the setting of college basketball. We find a large reduction, 34 percent, in the number of fouls committed during a basketball game when the number of referees increases from two to three. Additional empirical evidence is presented which suggests that this elastic supply of basketball crime is due to more competent officiating and cleaner play.

> I warne you wel, it is no childes pley. [GEOFFREY CHAUCER, *Canterbury Tales*]

I. Introduction

The economic approach to crime (Becker 1968, Tullock 1971, B[illegible] 1973; Ehrlich 1973, 1975; Becker and Landes 1974; Becker and Stigler 1974) develops the concept of a market for criminal activities. The supply of crime is a function of the costs and benefits of illegal activities and individual risk preference. The demand for crime derives from the free-lunch theorem—not all crime is worth preventing. In

---

We are indebted to the Clemson University Athletic Department, especially Bob Bradley, Bill Foster, and Bobby Robinson for their assistance on this project. Without implicating them, we are also grateful to Gary Becker, Rex Cottle, Marty Geisel, Michael Maloney, James Savarese, Edward Stertz, Earl Thompson, Gordon Tullock, and Lee Wakeman for helpful comments and interventions on our behalf.

[*Journal of Political Economy*, 1984, vol. 92, no. 2]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0004$01.50

this paper we employ these demand and supply functions to discuss the impact of the number of law enforcers on the arrest rate.

Our interest centers on the impact that a small change in the number of enforcers has on the number of arrests. Does a greater number of policemen lead to more or fewer arrests, other things equal?[1] To address this question we develop a theory that models the behavior of enforcers and criminals simultaneously. We motivate the discussion of the theory with an example from the world of sports. Our question in this context is, What is the impact of the addition of a third official in college basketball on the number of fouls called per game? Sports provide an economic laboratory with a history of accurate reporting of events. This is unlike most criminal data, which are generally held to be subject to various types of reporting error. We thus apply our theory in Section III by using the complete history (1954–83) of the Atlantic Coast Conference (ACC) basketball tournament. Employing data on fouls per game and controlling for a variety of ceteris paribus conditions (players' ability and experience, rule changes, attendance, and so forth), we find a statistically important *negative* association between the number of officials and the number of fouls called per game. Since this result can be due to several effects at work in our model, we present evidence in Section IV that the negative relation is due to better officiating and cleaner play. Concluding remarks are offered in Section V.

II. A Theory of Law Enforcement

In any given time period the probability that a criminal will commit a crime is a function of the expected costs and benefits of crime. The expected cost of crime is the probability of arrest and conviction times the fine associated with conviction; the expected benefit of crime is the probability of not being detected times the rewards of illegal behavior.

Formally,

$$P_C = P(D, F, B), \tag{1}$$

where $D = 1 - (1 - P_{A|C})^N$. P_C is the probability of criminal activity. D is the probability of detection and conviction given that a crime has been committed; $P_{A|C}$ is the probability that any one policeman will make an arrest given that a crime has been committed; F is the fine; N is the number of police; and B is the benefit of criminal activity. We assume that P_C decreases with an increase in D, an increase in F, and a reduction in B.[2]

---

[1] Swimmer (1974) addresses this problem in the context of seven different crimes. However, as he notes, there are several problems with his data. In particular, there are nonnormal errors in the reporting of crime statistics across cities and income groups by the offended person, and there is also nonnormality in the errors of police reports about crime statistics. We attempt to avoid such problems by applying our model to a data set that does not suffer from these problems.

The police can make an arrest whether or not a crime has been committed, and vice versa. Hence, we write

$$P_{A|C} = A(N) \tag{2}$$

and

$$P_{A|\bar{C}} = G(N), \tag{3}$$

where $P_{A|C}$ and N are defined as above and $P_{A|\bar{C}}$ is the probability of a false arrest. The signs of A_N and G_N are both ambiguous.

In the general case of N policemen, the probability of an arrest is

$$P_A = 1 - P_C(1 - P_{A|C})^N - (1 - P_C)(1 - P_{A|\bar{C}})^N. \tag{4}$$

Inserting the behavior of the actors into (4) yields

$$P_A = 1 - P(D, F, B)[1 - A(N)]^N - [1 - P(D, F, B)][1 - G(N)]^N, \tag{5}$$

with $D = 1 - [1 - A(N)]^N$, which is a reduced-form demand and supply theory of criminal activity.

We illustrate one comparative static result of the model, which addresses the central theme of the paper. Namely, the derivative of (5) with respect to N predicts the change in the number of arrests for an increase in the number of police, ceteris paribus. This derivative is:

$$\begin{aligned}\frac{\partial P_A}{\partial N} = {} & P_D D_N[(1 - G)^N - (1 - A)^N] + NPA_N(1 - A)^N \\ & + N(1 - P)G_N(1 - G)^{N-1} \\ & - P[(1 - A)^N \ln(1 - A)] \\ & - (1 - P)[(1 - G)^N \ln(1 - G)],\end{aligned}$$

where $D_N = NA_N(1 - A)^{N-1} - [\ln(1 - A)](1 - A)^N$.

We cannot predict the sign of (6). Increasing N increases th
ber of policemen who might detect a crime or make a false ar
last two terms are positive), but it changes the probability that
will do so (the second and third terms are ambiguous). *A fortiori*, criminal behavior adjusts to changes in the probability of arrest, the

---

[2] The individual's preference toward risk is crucial in mapping D, F, and B into P_C. For our purposes, however, the degree of risk aversion is not important. Whatever it is, we assume it is constant. Moreover, we interpret the fine associated with detected crime mean any sanction, including incarceration.

first term in (6). Even if we could sign the impact of N on police behavior—for example, that more police means better detection and more arrests—the response of criminals would be to commit fewer crimes, and the total effect remains ambiguous.[3] Even if more police reduces the probability that any one policeman will catch a criminal, the larger number of police might lead to an increased probability of detection as N increases. Consequently, it is not possible to say a priori whether more police is associated with more or fewer arrests.[4] Since the sign of $\partial P_A/\partial N$ cannot be determined logically, we seek an empirical estimate of its sign in the next section.

III. An Application to College Basketball

Our theory does not predict a sign for (6), that is, for what happens to the arrest rate when the number of police increases. To the extent that rules violations in basketball are analogous to criminal behavior, a data source is available that allows us to estimate the sign of (6) and to draw inferences about the impact of N on A and G. We recognize that the typical sanction meted to a basketball player who violates the rules is not the same thing as incarceration. However, many of the sanctions imposed for rules violations are parallel to putting someone in jail. For example, a player with five fouls goes out of the game. Two technical fouls on a coach lead to his ejection from the game. Strictly speaking, monetary fines are not imposed for rules violations (although they are in an opportunity cost sense). Rather, possession of the ball or a free throw is given to the other team. There is an empirical symmetry between basketball rules violations and crime, which means that an estimate of equation (6) in a basketball economy may be useful in predicting its sign in general.

In 1978, an experiment began in college basketball. The number of officials in the Atlantic Coast Conference (ACC) was increased from two to three per game. This change allows us to estimate the impact of more enforcers on the arrest rate, which is the number of personal fouls committed by each team in a game. Our data base consists of the history of the ACC tournament.[5] Beginning with the 1979 tournament, there were three officials assigned to each game. Prior to 1979, there were two. For most of the games we can identify the officials by name. In addition to the change in the number of officials, there have been other rule changes. In 1983, a 30-second shot clock and a 3-point shot from 19 feet were in effect. Starting in 1973, the first six fouls committed by either team in a half resulted in the other team gaining possession of the ball. Prior to 1973, a single free throw was taken by the fouled player. Beginning in 1963, an offensive foul cost a possession; prior to 1963, a fouled defensive player was awarded a free throw.

---

[3] Indeed, the first and second terms in (6) are probably of opposite sign. If A_N is positive, D_N is also positive, and the first term in (6) is negative ($A > G$). If increasing the size of the police force increases the individual policeman's probability of arrest (and if $A > G$), criminals respond by committing fewer crimes. The reverse is possible although not logically required.

[4] By similar argument it is also impossible to predict the impact of police competency on the number of arrests. Highly trained and experienced police forces may be associated with low arrest statistics. Moreover, in a profit-maximizing setting no owner would willingly bear the extra cost of adding an enforcer, such as shoplifting detectives in a department store, unless the crime rate would decrease. But as we have just noted, this does not imply more arrests. Decreases in the commission of crime do not necessarily reduce the arrest rate because more false arrests could be forthcoming as the number of police increases. The problem is compounded in nonproprietary settings because crime minimization may not be the objective function.

To estimate the sign of $\partial P_A/\partial N$, we also need to control for changes in F and B, the costs and benefits of fouling. We estimated the number of fouls called on the winning and losing team in each game as a function of the other team's field goal and free throw accuracy in the game, FGPCT and FREEPCT; the total score of the game, SCORE; the year of the tournament, TIME; the difference in the height of the winning and losing team, HITEDIFF; the difference in the experience of the winning and losing team measured by the number of lettermen, PLAYEXP; the difference in coaching experience measured by the number of years as coach at the respective schools, EXP; attendance at the game, ATTEND; and the experience of the referees as measured by the average number of ACC tournament games officiated by the crew, OFFEXP. SHOOT and CHARGE control the two rule changes discussed above. SHOOT is one for years all fouls were awarded a free throw and zero for years when th six fouls were penalized by loss of possession. CHARGE is ze years when an offensive foul cost the offending team possessi one for years when it cost a free throw. OFFICIAL controls t number of referees calling the game; it takes the value 2 or anticipate that SCORE controls for the 1983 rule changes. An appendix that gives our data and information about their derivation is available from the authors. The data were obtained from Barrier (1981) and various editions of the *Atlantic Coast Conference Basketball Yearbook*.

---

[5] The ACC holds its basketball tournament each March at the end of regular season play. Generally there have been eight member schools, although from 1972 to 1979 there were only seven schools. The University of North Carolina team did not participate in the 1961 tournament because it was guilty of a rule violation (playing an ineligible player). The teams are seeded according to regular season records, and the tournament is a single-elimination event. This means that eight teams yield a seven-game tournament and seven teams a six-game tournament. The first tournament was played in 1954, and we have data on the 1983 tournament. We are missing data for 1955 and 1962, and hence we have a total of 201 games in our data set

TABLE 1

REGRESSION RESULTS—BASKETBALL FOULS

	Parameter Estimate	Standard Error	t-Ratio	Prob > $\|t\|$
	Number of Fouls Committed by Winner (F-ratio = 14.90; R^2 = .5471)			
INTERCEPT	−4,709.49	602.680559	−7.8142	.0001
TIME	2.426270	.308124	7.8743	.0001
SCORE	.104337	.022676	4.6012	.0001
FREEPCT	−12.721683	4.687295	−2.7141	.0074
FGPCT	−14.601949	8.443828	−1.7293	.0858
SHOOT	11.107982	2.999642	3.7031	.0003
OFFICIAL	−21.059024	2.631219	−8.0035	.0001
HITEDIFF	.724399	.408193	1.7746	.0780
EXP	.238277	.085598	2.7837	.0061
PLAYEXP	−.398640	.247292	−1.6120	.1091
OFFEXP	−.428517	.152126	−2.8169	.0055
ATTEND	−.000145	.000397	−.3652	.7155
CHARGE	14.461917	3.670531	3.9400	.0001
	Number of Fouls Committed by Loser (F-ratio = 10.67; R^2 = .4639)			
INTERCEPT	−3,458.62	531.447697	−6.5079	.0001
TIME	1.778320	.271717	6.5447	.0001
SCORE	.124287	.023660	5.2531	.0001
FREEPCT	−3.950536	5.439805	−.7262	.4688
FGPCT	−22.704364	9.475005	−2.3962	.0178
SHOOT	9.922577	2.620516	3.7865	.0002
OFFICIAL	−13.784749	2.332086	−5.9109	.0001
HITEDIFF	−.138335	.358248	−.3861	.6999
EXP	.124235	.075083	1.6546	.1001
PLAYEXP	−.145511	.214533	−.6783	.4987
OFFEXP	−.280478	.134300	−2.0884	.0385
ATTEND	.000232	.000353	.6577	.5118
CHARGE	9.881283	3.360302	2.9406	.0038

The results of estimating this model by ordinary least squares are reported in table 1. In both the winner's and the loser's equation, OFFICIAL has a negative and statistically significant coefficient.[6] The magnitude of the coefficient suggests that adding the third official reduced the number of fouls called per game by about 17. Over the whole period, 1954–83, there were on average 52.2 fouls called per game. Hence, the effect of adding the third official was substantial—a 50 percent increase in the number of enforcers was associated with a 34 percent reduction in the number of arrests.

---

[6] We estimated the winner's and the loser's foul equation by the method of seemingly unrelated regressions. The results are comparable to those reported in table 1. Deleting the experience variables or ATTEND does not alter the results. We tested for equality of the OFFICIAL coefficients across the winner's and loser's equation in each of these specifications. In every case the impact of the third official is larger (more negative) on the winner than the loser. In the specification in table 1 the F-statistic is 4.28, which is significant at the 5 percent level. In the other two specifications, deleting the experience variables and deleting the experience variables plus ATTEND, the test suggests equality.

The other results in table 1 are consistent with rational behavior on the part of the participants. The time trend has been to increase the number of fouls called by about two per game per year. SCORE, as a proxy for speed and action on the court, is positively associated with more fouls per game; more action begets more fouls although the direction of effect could be the opposite. Each 10 points scored is associated with about one more foul. The better either team shoots, as proxied by FREEPCT and FGPCT, the less it is fouled. On average a 1 percentage point increase in free throw accuracy by one team is associated with about eight fewer fouls per game by the opponent. The greater the difference between the heights of the two teams, the more fouls per game. The more disparate is the experience of the two coaches in a given game, the more fouls are called. The opposite is true for players' experience. Referees' experience has a negative and significant coefficient. Attendance is a proxy for the importance of the outcome of the game.[7] The two rule changes, SHOOT and CHARGE, both have strong positive signs of about the same magnitude. This is interpreted to mean that players and coaches would rather give their opponents a free throw than possession of the ball.[8]

IV. Two versus Three Officials

The negative relationship between arrests and the number of cers in the case of basketball can be due to two things. More re are associated with either fewer false arrests, fewer criminal ac both. If the probability of a false arrest decreases with N, it is that the probability of a good arrest increases. Although it logically required that these two effects be of opposite sign, intuition suggests they are as an empirical matter. In addition, if the probability of arrest increases with the number of police, we expect that the probability of detection also increases. Consequently, we contend that the response of criminals to an increase in N will be to commit fewer crimes. Therefore, if the number of false arrests decreases with N, the negative sign on (6) must be due to two things—fewer false arrests and less crime committed. On the other hand, if the evidence indicates that the number of false arrests increases with N, the negative relationship must be due to fewer good arrests, a negative marginal product for the third official, or more referee shirking.

---

[7] Attendance is a poor proxy for the importance of the game for at least two reasons. We do not know ticket prices, and we cannot measure the extent of the television audience or revenues. Face value of tournament tickets is not very relevant since the schools employ two-part pricing schemes to allocate tournament tickets. We were unable to obtain a time series of these pricing schemes by schools.

[8] We tried several other control variables. These include whether there was an overtime, the halftime score, the difference in the final score, the difference in the halftime score, and the total number of field goals attempted per game. None of these variables are statistically important, nor did their inclusion affect the other results.

Upsets

Quantifying the number of false arrests is not an easy task. Therefore, we first attempt to sign G_N indirectly by examining the time series of upsets in ACC tournament competition. We argue that the fewer mistakes made by officials, the more likely the better team will win the game. Specifically, official error (Type I or Type II) increases the variance of the outcome of the game and decreases the probability that the better team will win. Assume that the number of officials is mean preserving with respect to the scoring of each team but that official error increases the variance of each team's scoring. It follows *that the mass of probability that the expected loser will win must* increase. We define an upset as a tournament game in which a lower-seeded team defeats a higher-seeded team. Tournament seeds are based on regular season records against conference opponents. If the quality of officiating increases with N, going from two to three officials reduces the number of upsets, and vice versa.

Consider the following model. The probability of an upset is a function of the costs and benefits of winning, the technical abilities of the teams, and luck. For example, young teams (in our data teams with few lettermen) may acquire relatively more basketball capital over a season than old teams. This implies that a young team will lose relatively more of its early conference games and enter the tournament with a low seed. By the end of the season the difference between the abilities of the old and the young team will have decreased. This will increase the probability of an upset, ceteris paribus. A similar human capital argument can be advanced for coaches. PLAYEXP and EXP control for the difference in player and coaching experience. To control for changes in the costs and benefits of winning, we also entered TIME, SCORE, SHOOT, and CHARGE. Finally, we included a variable, NCAA, which controls for changes in the rules for advancing to postseason play. Prior to 1975, only the winner of the ACC tournament advanced to the National Collegiate Athletic Association tournament. From 1975 to 1979, the tournament winner and one other team could advance. Since 1980, 4.25 teams have advanced on average. These rule changes reduce the incentive of the top-seeded teams to win the tournament; they are already assured a

TABLE 2

REGRESSION COEFFICIENTS—UPSET MODEL

OLS ESTIMATES (F-Ratio = 3 35; R^2 = .1310)

	Parameter Estimate	Standard Error	t-Ratio	Prob > $\|t\|$
INTERCEPT	−23 566823	27.882899	− 8452	3991
TIME	.012449	.014161	8790	.3806
SCORE	− 000456	001149	− 3975	6915
FXP	−.011492	004219	−2.7236	.0071
PLAYEXP	−.037984	011627	−3 2670	0013
SHOOT	.032936	141496	2328	8162
CHARGE	195689	.157978	1 2387	2171
OFFICIAL	−.355691	.192095	−1 8516	0657
NCAA	083083	076524	1.0857	2791

LOGISTIC REGRESSION ESTIMATES (MODEL χ^2 = 24 50, df = 8)

	Parameter Estimate	Standard Error	χ^2 Statistic	Prob > χ^2
INTERCEPT	−130 31612147	166 51167003	61	4338
TIME	06808087	08443313	65	.4201
SCORE	− 00235112	00658561	13	7211
EXP	−.06130248	02484204	6.09	.0136
PLAYEXP	.21896673	07135558	9.42	0022
SHOOT	42417153	86915017	24	6255
CHARGE	1 04057746	94075572	1 22	2687
OFFICIAL	−2 81048966	1 67979328	2 80	0943
NCAA	.76486641	.67016956	1 30	2537

place in the NCAA tournament based on their regular season p mance.

We report ordinary least squares and logistic regressions upset model in table 2, including OFFICIAL. The change fro to three officials is associated with a reduced probability of upse the result is statistically significant at the 10 percent level.[9] We pret this result to mean that the quality of officiating increased with the advent of three officials. That is, three officials reduced the variance in the outcome of tournament games, which implies that the quality of officiating increased with three referees. In turn this suggests that the negative sign of $\partial P_A/\partial N$ is due to fewer false arrests and reduced fouling.[10]

---

[9] The other results in the upset model are consistent with economic behavior For example, the NCAA rule changes decreased the incentive of top-seeded teams to win the tournament. Alternative specifications, e g., deleting changes in the rules of play or including the experience of officials, do not affect the results It may be true that differences in player and coach experience and ability do not perfectly control for the expected closeness of the game. For example, it would be nice to have betting odds or some other measure of pregame point spreads, but we were unable to locate such data for our time series.

Crime and Output

Transfers, including transfers due to crime, are associated with redistributive competition. Individuals compete to capture transfers and to resist having their wealth taken away. In other words, real resources are consumed in the transfer process. What is the implication of this argument for crime and output? As Tullock (1967, p. 231) says, "A successful bank robbery will inspire potential thieves to greater efforts, lead to the installation of improved protective equipment in other banks, and perhaps result in the hiring of additional policemen." This means that economies with high crime rates will be associated with low real production, ceteris paribus.

In basketball terms, Tullock's point can be put this way. Fouling but not being caught in the act will induce players to foul and to practice avoiding fouls or techniques to draw the attention of officials when they are fouled, and may lead to the addition of more referees. Such behavior reduces scoring because spending time in this way decreases a player's ability to perform other basketball feats such as scoring. For this reason we argue that expected errors in officiating induce players to learn and perform skills that *reduce* scoring. We can use this argument to estimate the signs of G_N and A_N. If the addition of the third official is associated with more scoring on the basketball court, we have another piece of evidence that three officials led to a lower crime rate in terms of the number of fouls actually committed during a game. If three officials are associated with less scoring output, we would infer that the basketball crime rate actually rose in the three-official era. A positive association between officials and output would imply that the crime rate declined.

---

[10] This result squares with the reasons given for the move to three referees by basketball intellectuals. For example, Professor Edward Steitz of Springfield College, Secretary-Editor of *NCAA Basketball Rules*, made such a case for three officials in a phone conversation with us. For an introduction to the art of officiating, including a technical discussion of two vs. three officials, see Bunn (1968, pp. 177–247). Alternative explanations for the move to the third official include attempts to improve the enforcement of rules other than fouling and the potential for wealth redistribution. The wealth redistribution argument will not stand empirical scrutiny on several grounds. The regular season and tournament winning percentages across schools during the two and three official era appear to be normally distributed. In addition, only two schools voted against the change to three officials. Finally, given that sports is an income-inelastic good in the quality dimension, we expect sports entrepreneurs to add officials and other monitoring technology over time as a means of reducing official error and nonproductive behavior by players. We note that during the drafting of the final version of this paper, the United States Football League announced that it will employ instant replays to reduce officials' errors in the 1984 season.

In a purely technical sense, output on a basketball court is the score of each team. This does not mean that the economic output of a basketball game is the score. Winning the game may be paramount to the fans. As a wise coach once observed, an ugly win is better than a pretty loss. Nonetheless, we believe that economic output and technical output are monotone transformations of each other.[11]

We can estimate the impact of more officials on scoring output in a basketball economy. Consider the following model of the winner's and the loser's score. TIME controls for trend. SHOOT and CHARGE control for rule changes. HITEDIFF, PLAYEXP, and EXP control for differences in players' and coaches' ability and experience across the two teams in a given game. FREEPCT and FGPCT control for game-specific performance of players. FOULS represents a form of victim compensation for detected crime. The more a team is fouled, the more victim compensation it receives in the form of possession of the ball or free throws and the higher its score on average. OFFICIAL is used to determine whether there is more or less real output with more police.

Table 3 reports ordinary least squares estimates of the model. The results are not sensitive to model specification.[12] The coefficient on OFFICIAL is positive in both equations, but it is only statistically significant in the winner's equation. There is therefore some reason to believe that three officials are associated with increased scoring output on the court. Further evidence on this point can be gained from the results on officials' experience. The coefficient on this variable is positive and significant in both equations. More experienced officials are associated with higher scores, presumably because they call
ter game. More referee learning means better refereeing, bett
ereeing means fewer fouls committed, and fewer fouls com
means more real output. This is additional evidence that more b
ball officials reduces the number of fouls committed and the n
of bad calls.

---

[11] One casual piece of evidence to support this contention is the prevalence of shot clocks in professional basketball. We also observe the emergence of a shot clock in college basketball as the scoring in games declined due to the increasing number of teams that resorted to stall-like play on offense. Viewed in this way, slowdown basketball, such as the four corners, is simply an extreme investment of basketball talent in a wasteful way.

[12] We also estimated the output model and the foul model simultaneously where each team's score and fouls are endogenous using the method of three-stage least squares. The estimates were not substantially different from the ordinary least squares estimates reported in tables 1 and 3.

TABLE 3

Regression Coefficients: Output Model

	Parameter Estimate	Standard Error	t-Ratio	Prob > $\|t\|$
	Winner's Score (F-ratio = 11.63; R^2 = .4853)			
INTERCEPT	3,010.916	1,023.48	2.9418	.0038
TIME	−1.540839	.524111	−2.9399	.0038
OPPONENT'S FOULS	.713351	.130790	5.4542	.0001
CHARGE	.610939	5.984444	.1021	.9188
SHOOT	−10.106034	4.648230	−2.1742	.0313
OFFICIAL	7.286445	4.485530	1.6244	.1064
HITEDIFF	1.027700	.619922	1.6578	.0995
EXP	.037159	.132085	.2813	.7789
PLAYEXP	.205872	.371958	.5535	.5808
ATTEND	−.000745	.000607	−1.2263	.2221
OFFEXP	.539729	.231331	2.3331	.0210
FREEPCT	22.587973	9.298198	2.4293	.0163
FGPCT	126.528813	13.920502	9.0894	.0001
	Loser's Score (F-ratio = 4.64, R^2 = .2734)			
INTERCEPT	2,569.971	1,095.331	2.3463	.0203
TIME	−1.293030	.561771	−2.3017	.0227
OPPONENT'S FOULS	.524618	.120858	4.3408	.0001
CHARGE	−1.912368	6.038176	−.3167	.7519
SHOOT	−14.397993	4.787349	−3.0075	.0031
OFFICIAL	2.023766	4.921747	.4112	.6815
HITEDIFF	−.397539	.648689	−.6128	.5409
EXP	−.137062	.138543	.9893	.3241
PLAYEXP	−.222621	.392002	−.5679	.5710
ATTEND	−.000392	.000625	−.6270	.5316
OFFEXP	.405802	.242835	1.6711	.0968
FREEPCT	22.806112	7.402350	3.0809	.0025
FGPCT	45.123960	13.307591	3.3908	.0009

V. Conclusion

We present a model of criminal and police behavior where the behavior of each is endogenous. One important result of the model is that the arrest rate may be a misleading statistic when attempting to evaluate the quality and quantity of law enforcement. The mass of the sports evidence suggests that increasing the size of the police force reduces the arrest rate and increases the quality of law enforcement. No policy implications follow from this result, however. Increasing the size of the police force is not free, and we are not prepared to argue that such changes would be cost effective in practice. We can note, however, that the supply of crime in our data is relatively elastic. We find that a 50 percent increase in the number of officials is associated with a 34 percent reduction in arrests. Our other evidence suggests that the crime rate went down even more than this. We find that referee competency went up with the move to three officials and that the number of fouls called went down. So the crime rate must have decreased by more than 34 percent. This implies an elastic response by criminals to changes in the probability of being arrested.

References

Barrier, Smith. *The ACC Basketball Tournament Classic* Burlington, N.C : Metro Sports, 1981.

Barro, Robert J. "The Control of Politicians: An Economic Model." *Public Choice* 14 (Spring 1973): 19–42.

Becker, Gary S. "Crime and Punishment: An Economic Approach." *J P.E* 76 (March/April 1968): 169–217.

Becker, Gary S., and Landes, William M., eds. *Essays in the Economics of Crime and Punishment.* New York: Columbia Univ. Press (for N B E.R.), 1974.

Becker, Gary S., and Stigler, George J "Law Enforcement, Malfeasance, and the Compensation of Enforcers." *J. Legal Studies* 3 (January 1974): 1–18.

Bunn, John W. *The Art of Officiating Sports.* Englewood Cliffs, N.J : Prentice-Hall, 1968.

Ehrlich, Isaac. "Participation in Illegitimate Activities: A Theoretical and Empirical Investigation." *J.P.E.* 81 (May/June 1973). 521–65.

———. "The Deterrent Effect of Capital Punishment: A Question of Life and Death." *A.E.R.* 65 (June 1975): 397–417.

Office of the Commissioner and the Service Bureau. *Atlantic Coast Conference Basketball Yearbook.* Greensboro, N.C.: Atlantic Coast Conference. 1954–83. various editions.

Swimmer, Eugene. "Measurement of the Effectiveness of Urban Law Enforcement—a Simultaneous Approach." *Southern Econ. J* 40 (April 1974): 618–30.

Tullock, Gordon. "The Welfare Costs of Tariffs, Monopolies *Western Econ. J* 5 (June 1967): 224–32

———. *The Logic of the Law.* New York: Basic, 1971

The Effects of Increased Copyright Protection: An Analytic Approach

Ian E. Novos

University of Pennsylvania

Michael Waldman

University of California, Los Angeles

Previous authors who have considered partially nonexcludable goods have claimed that an increase in copyright protection will have the following two effects on social welfare. First, it will decrease the social welfare loss due to underproduction. Second, it will increase the social welfare loss due to underutilization. In this paper we investigate these claims in a formal setting by analyzing a model in which consumers vary only in terms of their costs of obtaining a reproduction. Our analysis provides partial support to the first claim of these previous authors while giving little or no support to the second claim

There are many commodities other than public goods that are characterized by aspects of nonexcludability. For example, any recording is partially nonexcludable if copies of the recording can be made by

We would like to thank Beth Allen, William Baumol, Jere Behrman, Doug Blair, David Cass, Dan Friedman, Henry Glick, Robert Inman, Jack Hirshleifer, Duncan Mann, Robert Margo, Michael Riordan, William Wascher, an anonymous referee, and George Stigler for comments on earlier drafts, and the Sloan Foundation and the Foundation for Research in Economics and Education for financial support. Special thanks go to Oliver Williamson for comments received and for conversations that helped lead us to the topic of this paper, and to John Riley for very helpful comments on a previous draft. We, of course, accept responsibility for any remaining errors.

[*Journal of Political Economy*, 1984, vol. 92, no. 2]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0008$01.50

individuals other than the original producer.[1] In this paper we look at markets of this type, paying special attention to the effects of copyright protection.[2]

Consider, for instance, the production and marketing of a computer software package in the form of computer disks. Given that the technology for copying computer disks is available to the public, some agents will be able to get access to the software package without paying the producing agent. This potential for free riding implies that the quality of the package provided via the market will be below the socially optimal quality. This is what has come to be known in the literature as the social welfare loss due to underproduction. This loss is analogous to the loss that occurs in the classic public goods case (see, e.g., Samuelson 1954). Unlike free riding in the case of a pure public good, however, free riding here entails the expenditure of resources to gain access to the software package (e.g., one has to borrow the package from some other agent, gain access to the copying technology, and purchase blank disks). This, combined with the fact that the producer's price will typically exceed his marginal cost of production because of the market power he possesses, leads to a second social welfare loss that is known as the social welfare loss due to underutilization. This social welfare loss has two components. The first is the loss due to consumers who would be willing to pay the marginal cost of production but who do not consume the good. The second is the loss due to consumers who expend more real resources in copying than would be incurred by the producer if these consumers purchased the software package from the producer.

Economists who have previously considered this ty
have claimed that an increase in copyright protection
following two effects on social welfare. First, it will dec
welfare loss due to underproduction. Second, it will inc
welfare loss due to underutilization.[3] In this paper

---

[1] We can provide a precise definition for our term "partial nonexclud
contrasting it with the concept of pure nonexcludability. Pure nonexcludability occurs when one individual cannot get access to a commodity without every other individual in the economy getting free access to the commodity. Partial nonexcludability occurs when the situation deviates from the preceding in one or more of the following ways (1) When one individual gets access to a commodity only a segment of the rest of the population gets access to that commodity. (2) When one individual gets access to a commodity other individuals are only afforded access to the commodity at some cost. (3) When one individual gets access to a commodity other individuals are only afforded access to a related commodity

[2] Note that, although we emphasize copyright protection, our results hold for any government policy that tends to limit nonexcludability. One example of such a noncopyright barrier is a tax placed on equipment consumers utilize to make reproductions, a proposal currently being debated in the U S. Congress

[3] See Hirshleifer and Riley (1979) for a clear statement of these claims, and Arrow

these claims in a formal setting by analyzing a model in which consumers differ only in terms of their costs of obtaining a reproduction. We first demonstrate that, given a restriction on the distribution of reproduction costs in the population, an increase in copyright protection does indeed decrease the social welfare loss due to underproduction. However, we then show through the analysis of a specific example that, when this restriction is not satisfied, an increase in copyright protection can actually lead to an increase in the social welfare loss due to underproduction. Finally, in contrast to the partial support we give to the first claim of previous authors, we provide little or no support for the second. That is, we derive a proposition more consistent with the social welfare loss due to underutilization decreasing, rather than increasing, when copyright protection is increased. The reason our results on this aspect of the issue so strongly contradict the claims of previous authors is that our consumers do not vary in terms of their valuations of the partially nonexcludable good. We conjecture, however, that even in a world where consumers differ in this way, it would still be the case that an increase in copyright protection could lead to a decrease in the social welfare loss due to underutilization.

The outline for this paper is as follows. Section I sets forth a model of an economy containing a partially nonexcludable good. Sections II and III analyze this model, paying special attention to the social welfare consequences of the government varying the extent of copyright protection. Section IV presents some concluding remarks.

I. The Model

We consider a market wherein a partially nonexcludable good, denoted good M, is produced by a monopolist. M is indivisible—in particular, each individual will consume either zero units or one unit of M. Equation (1) describes the monopolist's total costs, denoted $TC(x, Q)$, of producing x units of M of quality Q:

$$TC(x, Q) = F(Q) + cx, \tag{1}$$

where $F(Q)$ denotes the fixed costs of production, and where cx denotes the variable costs of production. The function $F(Q)$ is also assumed to satisfy the following restrictions: $F(0) = 0$, $F'(\cdot) > 0$, and $F'(\infty) = \infty$.

Individuals derive a return from the consumption of M that depends on the quality of the unit consumed, where the quality of the

---

(1962) for a related discussion. See also Ploman and Hamilton (1980) for a discussion by noneconomists in which similar claims are made.

unit consumed by individual i is denoted Q_i. There are two ways that a consumer can acquire a unit of M. The first method is for the consumer to purchase M directly from the monopolist. The second method is for the consumer to borrow a unit from some other agent who has in one way or another acquired it and then make a reproduction. We refer to the monopolist as the primary source for obtaining M and to agents who lend out M as secondary sources. Agents who lend out units of M incur no costs in the lending process. If consumer i acquires a unit of M by borrowing a unit from a secondary source and copying it—a process we will refer to as going through the secondary market—then he incurs reproduction costs $c + z_i(1 + H)$, where H denotes the extent of copyright protection.[1] One interpretation for the term $z_i(1 + H)$ is that it is the cost the consumer incurs, which is due to the time he expends in making a reproduction. Note that the specification above ensures that the marginal cost of production for the monopolist is less than any consumer's private cost for reproducing the monopolist's output. That is, to produce an extra unit of output the monopolist incurs a cost c, while to make a reproduction each consumer i incurs a cost $c + z_i(1 + H)$.

The following profit function, which incorporates the properties just discussed, is adopted:[5]

$$\Pi_i = vQ_i - e_i. \tag{2}$$

In specifying (2) we have assumed that each consumer has the same valuation for quality, denoted v. Additionally, e_i denotes the costs individual i incurs in acquiring good M. Note also that if individual i does not consume a unit of M, $Q_i = 0$.

As stated previously, we assume that individuals differ in te
their costs of obtaining a reproduction. In terms of ou
means that the z_i's are not constant across individuals. K
distributed between the extreme values 0 and Z, where
tion is described by a density function $g(\cdot)$ that is continu
entiable and nonzero in the specified interval. We also
each consumer i's value for z is unobservable to everyone
and that there exists a Q such that $\int_{vQ-c}^{Z}(vQ - c)g(z_i)dz_i$ ... $(Q$
last restriction guarantees that the monopolist always has an incentive to produce.

---

[1] Alternative ways to incorporate copyright protection would be to have the costs of going through the secondary market equal either $c + z_i + H$ or $(c + z_i)(1 + H)$. We have analyzed the model under these alternative specifications and found that the results do not change significantly.

[5] A seemingly more general specification than the one employed in (2) would be to have the return derived from the consumption of M be a general function $v(Q)$. In actuality, however, this yields little additional generality since it would always be possible to rescale the quality variable to yield the linear form used in this paper

We will end this section by deriving a pair of equations that characterize the socially optimal quality of M. To do this we assume for the moment that the technology for producing M is in the hands of a social-welfare-maximizing government, which produces the good and makes units of the good available to consumers at marginal cost. (Note: This government is assumed to finance any expenditures through lump-sum taxes.) We also assume that the government's social welfare function is simply the sum of the profits of all the individuals in the economy. Thus, the socially optimal quality of M, denoted Q^*, is defined by equation (3):[6]

$$Q^* = \arg\max_{Q} \int_0^Z (vQ - c)g(z_i)dz_i - F(Q). \tag{3}$$

The maximization problem in (3), in turn, yields the following first-order condition:

$$\int_0^Z vg(z_i)dz_i - F'(Q) = 0. \tag{4}$$

II. The Social Welfare Loss Due to Underproduction

Before proceeding to analyze the social welfare loss due to underproduction, a number of points need to be made. First, the monopolist will only offer a single quality of M for sale. This is a consequence of each consumer having the same valuation for quality. That is, there is no opportunity for the monopolist to price discriminate by offering a variety of qualities, because all consumers purchasing from the monopolist have the same valuation for quality.[7] Second, because secondary sources incur no costs in the lending process, in equilibrium they will lend the same quality good the monopolist is selling, but at no charge to the borrower. This last point tells us which consumers will prefer to borrow rather than purchase. A consumer will prefer to borrow rather than purchase when his costs of acquiring the good through the secondary market are less than the price being charged by the monopolist, denoted P. Given that secondary sources lend the good at no charge, this implies that consumer i will prefer to borrow rather than purchase when $c + z_i(1 + H) < P$, or $z_i < (P - c)/(1 + H)$.

In the previous section we derived a pair of equations that characterize the socially optimal quality of M. We will now compare that quality with the quality actually produced by the monopolist.

---

[6] This is a correct method of finding the socially optimal quality of M because in our equilibrium all individuals, through either the primary market or the secondary market, acquire a unit of M.

[7] A formal proof of this point is available from the authors upon request.

Proposition 1: $Q^M(H) < Q^*$ for all H, where $Q^M(H)$ denotes the quality of M produced by the monopolist.

Proof: In choosing a (P, Q) pair, the monopolist has to take into account the following two facts. First, no consumer would be willing to pay for M an amount greater than his valuation of M; that is, vQ must exceed or equal P. Second, each consumer for whom $z_i < (P - c)/(1 + H)$ will prefer to go through the secondary market rather than purchase from the monopolist. Together these two facts imply that the monopolist's maximization problem is the following:

$$\max_{P,Q} \int_{(P-c)/(1+H)}^{Z} (P - c)g(z_i)dz_i - F(Q), \tag{5}$$

subject to $vQ \geq P$. It is easy to demonstrate that the constraint $vQ \geq P$ must hold as an equality. Therefore, $Q^M(H)$ is defined by the following equation:

$$Q^M(H) = \arg\max_{Q} \int_{S(Q,H)}^{Z} (vQ - c)g(z_i)dz_i - F(Q), \tag{6}$$

where $S(Q, H) = (vQ - c)/(1 + H)$. The maximization problem in (6), in turn, yields the following first-order condition:

$$\int_{S(Q,H)}^{Z} vg(z_i)dz_i - v(1 + H)^{-1}(vQ - c)g[S(Q, H)] - F'(Q) = 0. \tag{7}$$

A comparison of (4) and (7) yields $Q^M(H) \neq Q^*$ for all H. Furthermore, combining this with equations (3) and (6) yields th and (9) hold for all H:

$$\int_0^Z [vQ^M(H) - c]g(z_i)dz_i - F[Q^M(H)]$$
$$< \int_0^Z (vQ^* - c)g(z_i)dz_i - F(Q^*),$$

$$\begin{aligned} &\int_{S[Q^M(H),H]}^{Z} [vQ^M(H) - c]g(z_i)dz_i - F[Q^M(H)] \\ &\quad > \int_{S(Q^*,H)}^{Z} (vQ^* - c)g(z_i)dz_i - F(Q^*). \end{aligned} \tag{9}$$

Equations (8) and (9), in turn, imply that (10) holds for all H.

$$\int_0^{S[Q^M(H),H]} [vQ^M(H) - c]g(z_i)dz_i < \int_0^{S(Q^*,H)} (vQ^* - c)g(z_i)dz_i. \tag{10}$$

Finally, (10) yields $Q^M(H) < Q^*$ for all H. Q.E.D.

Proposition 1 demonstrates that there is always a social welfare loss due to underproduction.[8] Our next step is to investigate how a change in H affects this social welfare loss.

PROPOSITION 2: If $g'(z_i) \geqslant 0$ for all $0 < z_i < Z$, then $Q^M(H_2) > Q^M(H_1)$ for any H_1, H_2 pair, where $H_2 > H_1$.

PROOF: Consider an H_1, H_2 pair, where $H_2 > H_1$. Equation (6) yields both (11) and (12):

$$\begin{aligned}&\int_{S[Q^M(H_1),H_1]}^{Z}[vQ^M(H_1) - c]g(z_i)dz_i - F[Q^M(H_1)]\\ &\geqslant \int_{S[Q^M(H_2),H_1]}^{Z}[vQ^M(H_2) - c]g(z_i)dz_i - F[Q^M(H_2)],\end{aligned} \tag{11}$$

$$\begin{aligned}&\int_{S[Q^M(H_2),H_2]}^{Z}[vQ^M(H_2) - c]g(z_i)dz_i - F[Q^M(H_2)]\\ &\geqslant \int_{S[Q^M(H_1),H_2]}^{Z}[vQ^M(H_1) - c]g(z_i)dz_i - F[Q^M(H_1)].\end{aligned} \tag{12}$$

Equations (11) and (12), in turn, yield equation (13):

$$\begin{aligned}&\int_{S[Q^M(H_1),H_2]}^{S[Q^M(H_1),H_1]}[vQ^M(H_1) - c]g(z_i)dz_i\\ &\leqslant \int_{S[Q^M(H_2),H_2]}^{S[Q^M(H_2),H_1]}[vQ^M(H_2) - c]g(z_i)dz_i.\end{aligned} \tag{13}$$

Furthermore, given our restriction on the density function $g(\cdot)$, (13) implies $Q^M(H_2) \geqslant Q^M(H_1)$.

Now, given the result above, $Q^M(H_2) = Q^M(H_1)$ can only occur if at H_1 the partial derivative of the left-hand side of (7) with respect to H is equal to zero. Taking this derivative, setting it equal to zero, and rearranging yields

$$2g\{S[Q^M(H_1), H_1]\} = -S[Q^M(H_1), H_1]g'\{S[Q^M(H_1), H_1]\}. \tag{14}$$

Equation (14) violates our restriction on $g(\cdot)$. Q.E.D.

Proposition 2 demonstrates that, given a restriction on the density function $g(\cdot)$,[9] our model displays a property consistent with the first claim of economists who have previously considered this type of market. That is, given this restriction, an increase in copyright protection necessarily results in a decrease in the social welfare loss due to underproduction. When this restriction is not satisfied, however, it is possible for an increase in copyright protection to actually lead to an increase in the social welfare loss due to underproduction. This can be seen through the following example. Let $Z = 1, v = 1, c = 0, F(Q) = Q^2, g(z_i) = 6$ for all $0 \leqslant z_i \leqslant \frac{1}{4}$ and $g(z_i) = 3$ for all $\frac{1}{4} < z_i \leqslant 1$. Analysis of this example yields that $Q^M(1) = \frac{3}{5}$ and $Q^M(\frac{1}{2}) \cong .55$.

---

[8] Proposition 1 holds for all finite values of H. When $H = \infty$, (4) and (7) are equivalent. This simply means that when the good is fully excludable, there is no social welfare loss due to underproduction.

[9] One density function that satisfies this restriction is the uniform density function.

One might ask why, in contrast to the claims of previous authors, the social welfare loss due to underproduction does not always decrease when copyright protection is increased. The logic behind the claims of previous authors is as follows. As copyright protection is increased, more individuals will be forced to purchase from the monopolist. Thus, the additional revenue that will result from a marginal increase in quality will be higher, the greater is the extent of copyright protection. Furthermore, this in turn implies that the greater is the extent of copyright protection, the higher will be the quality of good produced by the monopolist and the lower will be the social welfare loss due to underproduction.

Now consider equation (7). This equation tells us that an increase in copyright protection actually affects the monopolist's choice of quality through two channels. The first is the one recognized by previous authors. That is, at any given quality level, the number of individuals who purchase from the monopolist will increase when copyright protection is increased. As argued by previous authors, this channel tends to make $Q^M(\cdot)$ a positive function of H. The second channel has to do with the idea that as Q is increased, some consumers shift from the primary market to the secondary market. The shifting of consumers from the primary to the secondary market exerts a negative
fluence on Q, the magnitude of this influence being posit
to the number of people who shift. When $g(\cdot)$ is a n
function, the number of people who shift at any given
be decreasing with H. This in turn implies that as H i
severity of this negative influence will decrease. This
proposition 2. That is, when $g(\cdot)$ is a nondecreasing fu
channels work in the same direction and cause $Q^M(\cdot)$ to be a
function of H. When $g(\cdot)$ is a decreasing function, however, this second channel may work in the direction of making $Q^M(\cdot)$ a negative function of H. Thus, when this is the case, the intuition of previous authors may be incorrect because a factor they did not recognize can push $Q^M(\cdot)$ in the direction opposite to what they expected.[10]

---

[10] One might wonder whether the results concerning $Q^M(\ )$ in our example derive from the discontinuous nature of $g(\cdot)$, as opposed to the decreasing nature of $g(\ )$. Given that the following can easily be demonstrated, our results concerning $Q^M(\ )$ in this example must derive from the decreasing nature of $g(\cdot)$. Corollary to proposition 2. Assuming nothing concerning the continuity of $g(\ )$, if $g(z_2) \geqslant g(z_1)$ for any z_1, z_2 pair, $0 \leqslant z_1 < z_2 \leqslant Z$, then $Q^M(H_2) \geqslant Q^M(H_1)$ for any H_1, H_2 pair, $H_2 > H_1$.

III. The Social Welfare Loss Due to Underutilization

In the previous section we investigated how a change in H affects the social welfare loss due to underproduction. We will now investigate how a change in H affects the social welfare loss due to underutilization.

PROPOSITION 3: For every H_1, there exists an H_2, $H_2 > H_1$, such that $W(H_3) < W(H_1)$ for every $H_3 \geqslant H_2$. Note that $W(H)$ denotes the social welfare loss due to underutilization.

PROOF: In our model, all consumers obtain M in equilibrium. Thus, the social welfare loss due to underutilization is simply the costs incurred in the secondary market that are in excess of what would be incurred if the good were purchased from the monopolist. Equation (15) captures this formally:

$$W(H) = \int_0^{S[Q^M(H), H]} z_i(1 + H)g(z_i)dz_i. \tag{15}$$

Now consider some random value for H, denoted H_1. We will denote as H_2 the value for H at which (16) holds:

$$\int_0^{S(Q^*, H_2)} (vQ^* - c)g(z_i)dz_i = W(H_1). \tag{16}$$

Proposition 1, together with the definition of $W(H)$, yields (17):

$$W(H_3) < \int_0^{S(Q^*, H_2)} (vQ^* - c)g(z_i)dz_i \quad \text{for all } H_3 \geqslant H_2. \tag{17}$$

Equations (16) and (17) together, in turn, yield $W(H_3) < W(H_1)$ for all $H_3 \geqslant H_2$. Q.E.D.

Proposition 3 demonstrates that our model does not support the prevailing orthodoxy concerning how the social welfare loss due to underutilization behaves as copyright protection is increased. That is, there seems to be no tendency for an increase in copyright protection to increase the social welfare loss due to underutilization.

As with our previous result, one might ask why we reach a different conclusion than previous authors do. One difference between our analysis and that of previous authors is that they did not take into account the costs involved in obtaining a partially nonexcludable good through a secondary market. The reason such costs are important is the following. Since the monopolist is pricing above marginal cost, consumers at the margin are expending more resources in acquiring the good through the secondary market than would be incurred if the good were purchased from the monopolist. Thus, when individuals shift from the secondary to the primary market as a result of an increase in copyright protection, the social welfare loss due to underutilization tends to decrease.

The discussion above brings to the fore one other interesting point. There is a widely recognized social welfare loss that occurs when economic agents expend resources surmounting or evading barriers. For example, the social welfare loss attributable to the resources expended in tariff evasion has been analyzed by, among others, Johnson (1972), Bhagwati and Hansen (1973), and Bhagwati and Srinivasan (1973); while Tullock (1967) argues that a major portion of the social welfare loss associated with theft derives from the resources agents invest in the activity of theft. All of these previous papers, however, have been concerned with activities that "yield pecuniary returns but do not produce goods or services that enter a utility function directly or indirectly via increased production or availability to the economy of goods that enter a utility function" (Bhagwati 1982, p. 989). Our analysis shows that this type of social welfare loss can arise even when the activity under consideration involves positive production, as making a reproduction surely does.

IV. Limitations and Extensions

Previous authors who have considered partially nonexcludable goods have claimed that an increase in copyright protection will lead to a decrease in the social welfare loss due to underproduction and to an increase in the social welfare loss due to underutilization. In this paper we have found partial support for the first claim but little or no support for the second. The reason our results concerning the social welfare loss due to underutilization so strongly contradict the [illegible] previous authors is that our consumers did not vary [illegible] tions on the quality of the partially nonexcludable good [illegible] tion is crucial because, if copyright protection were to [illegible] a world where this assumption did not hold, some [illegible] stop consuming the partially nonexcludable good [illegible] ture, however, that even in such a world the basic point [illegible] demonstrate would be valid. That is, it would still be the [illegible] increase in copyright protection *could* lead to a decrease in the social welfare loss due to underutilization.

Finally, we would like to end this paper by suggesting two different ways that our analysis could be extended. First, consumers could be allowed to vary in terms of their valuations on the quality of the partially nonexcludable good. With this type of variation the monopolist would need to be concerned with self-selection constraints, an issue central to a number of recent papers (e.g., Spence 1977; Mussa and Rosen 1978; Chiang and Spatt 1982; Maskin and Riley 1982). Second, in the real world it is not only the case that consumers expend resources when they obtain a partially nonexcludable good from someone other than the original producer, but also that producers expend resources in an attempt to make this type of acquisition more difficult. Thus, a second worthwhile extension might be to incorporate these types of expenditures into our model.

References

Arrow, Kenneth J. "Economic Welfare and the Allocation of Resources for Invention." In *The Rate and Direction of Inventive Activity: Economic and Social Factors* Universities–National Bureau Conference Series. Princeton, N.J.: Princeton Univ. Press (for N.B.E.R.), 1962.

Bhagwati, Jagdish N. "Directly Unproductive, Profit-seeking (DUP) Activities." *J.P.E.* 90 (October 1982): 988–1002.

Bhagwati, Jagdish N., and Hansen, Bent. "A Theoretical Analysis of Smuggling." *Q.J.E.* 87 (May 1973). 172–87.

Bhagwati, Jagdish N., and Srinivasan, T. N. "Smuggling and Trade Policy." *J Public Econ.* 2 (November 1973): 377–89.

Chiang, Raymond, and Spatt, Chester S. "Imperfect Price Discrimination and Welfare." *Rev. Econ. Studies* 49 (April 1982): 155–81.

Hirshleifer, Jack, and Riley, John G. "The Analytics of Uncertainty and Information—an Expository Survey." *J. Econ. Literature* 17 (December 1979): 1375–1421.

Johnson, Harry G. "Notes on the Economic Theory of Smuggling." *Malayan Econ. Rev.* 17 (April 1972): 1–7. Reprinted in *Illegal Transactions in International Trade: Theory and Measurement*, edited by Jagdish N. Bhagwati. Amsterdam: North-Holland, 1974.

Maskin, Eric S., and Riley, John G. "Monopoly Selling Strategies with Incomplete Information." Working Paper no. 268, UCLA, October 1982.

Mussa, Michael, and Rosen, Sherwin. "Monopoly and Product Quality." *J. Econ. Theory* 18 (August 1978): 301–17.

Ploman, Edward W., and Hamilton, L. Clark. *Copyright. Intellectual Property in the Information Age.* Boston: Routledge & Kegan Paul, 1980.

Samuelson, Paul A. "The Pure Theory of Public Expenditure." *Rev Econ and Statis.* 36 (November 1954): 387–89.

Spence, A. Michael. "Nonlinear Prices and Welfare." *J. Public Econ.* 8 (August 1977): 1–18.

Tullock, Gordon "The Welfare Costs of Tariffs, Monopolies, and Theft." *Western Econ. J* 5 (June 1967): 224–32.

Conversion to a Consumption Tax: The Transition in a Life-Cycle Growth Model

Laurence S. Seidman

University of Delaware

This paper examines the transitional losses that may follow conversion from an income tax to a consumption tax, even when conversion eventually raises steady-state lifetime welfare. Two recent studies have left the impression that the transitional losses would be large. Using a life-cycle growth model, this paper examines two reasons why these studies have probably overestimated the transitional losses. The model computes the welfare gains and losses of each age cohort along the transition path. It is shown that the greater the bequest motive and subjective discount rate generating the particular initial income tax steady state, the smaller would be the transitional losses. Similarly, a policy of age phasing would reduce the losses. The results demonstrate the importance of examining the transition path whenever a policy change alters capital accumulation

The merit of converting the income (Y) tax to a pers
tion (C) tax has long engaged the interest of economis
previous studies concentrate exclusively on steady stat
studies recognize the importance of examining the t
(Summers 1979; Auerbach and Kotlikoff 1983). Simul
sion in a life-cycle growth model, these researchers find
significant fraction of cohorts alive at the time of conversion—in particular, those older than some age X—would lose significantly, even

---

I am grateful to my research assistants, Ruth Haas and Larry Ehmer, for writing the computer programs, implementing the computer calculations and simulations, and checking mathematical derivations and formulas. I am also grateful to the National Science Foundation for partial funding of research assistance

[*Journal of Political Economy*, 1984, vol. 92, no. 2]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0002$01.50

when there is ultimately a substantial steady-state gain in lifetime utility that will be enjoyed by all future cohorts. These studies, therefore, may give the impression that conversion to a *C* tax would cause a substantial transitional loss.

Why do cohorts older at the time of conversion lose significantly in these two studies? In the initial income tax steady state, each person's tax path over his life parallels his income path, which rises until retirement but then declines until death. In contrast, on conversion to a consumption tax, each person's new tax path parallels his consumption path, which, according to the life-cycle model, grows at a constant rate over the person's entire life, including retirement. Conversion to a *C* tax that raises the same total revenue from the whole population, therefore, will usually significantly raise the tax burden on older cohorts.

If a person's consumption path, contrary to the life-cycle model, ran "parallel" to his income path—rising during work years, but then declining with retirement—then conversion would not impose a significant loss on older cohorts. This study, however, will follow the two previous studies in using the life-cycle model. If consumption does decline with retirement, then all three studies exaggerate the transitional loss to older cohorts.

This study will focus on two other reasons why the Summers and Auerbach-Kotlikoff simulations may overstate the severity of the transitional problem under actual conversion: (1) the absence of a bequest motive and (2) the absence of a policy of age phasing.

The strict life-cycle model used by both studies excludes bequests. Yet recent empirical work suggests that a bequest motive may be responsible for an important share of actual capital accumulation (Brittain 1978; White 1978; Mirer 1979; Kotlikoff and Summers 1981). It should be noted that, in this paper, the phrase "bequest motive" does not necessarily imply that a person is concerned about the welfare of his heirs. The phrase is intended to include any reasons why a person might choose a consumption path that results in a bequest at death. For example, the desire to accumulate wealth for security, prestige, ego gratification, and as insurance against living "too long" might all account for a bequest motive.

Why is a bequest motive important for our transitional loss problem? The key point is that alternative combinations of bequest motive and subjective discount rate can generate the same steady state (characterized by a particular capital-effective labor ratio, saving rate, and interest rate). The higher the bequest motive, the higher would be the saving rate; but the higher the subjective discount rate, the lower would be the saving rate. Hence, by raising both appropriately, the steady-state saving rate can be held constant.

Suppose that under a Y tax the economy is observed to be in a neoclassical growth model steady state with a particular saving rate. This steady state may have been generated by a zero bequest motive and a particular discount rate, or by a positive bequest motive and a higher discount rate. The stronger the bequest motive, the higher the discount rate that must have generated the initial steady state.

The higher the discount rate, the slower is the growth rate of the consumption path over a person's life. The lifetime consumption path would be flatter, and consumption during retirement would be less. Conversion from a Y tax to a C tax would therefore be less harmful to retirees. There would be less divergence between the lifetime income tax path, which drops at retirement, and the lifetime consumption tax path, which now would grow more slowly and have a lower average value through retirement. Conversion would therefore impose a smaller loss on older cohorts.

Hence, it is the magnitude of the subjective discount rate, and the resulting flatness of the lifetime consumption path—and consumption tax path—that is crucial. By omitting a bequest motive, the two studies cited above assume that a unique subjective discount rate generates a particular initial steady state. But once a bequest motive is included, the stronger the bequest motive, the higher the subjective discount rate associated with a particular initial steady state. The higher the discount rate, the lower are the transitional losses following conversion.

The second reason the two studies may overstate the severity of the transitional problem under actual conversion is that they do not consider a policy of age phasing. One method of age phasin [illegible] all cohorts immediately under the C tax on conversio [illegible] tax rates for older cohorts. A second method would [illegible] older than some age D under the Y tax for the rest of [illegible] place all cohorts D or less permanently under the [illegible] the whole population would be under the C tax.

To isolate the importance of these two factors, this study [illegible] similar life-cycle growth model, except that it will include bequests and inheritances and will examine a policy of age phasing.

Section I describes the steady state of the life-cycle growth model utilized in this study, Section II the simulation model that generates the path of the economy under conversion, Section III conversion in a no-bequest economy, Section IV conversion in an economy with a bequest motive, and Section V conversion with age phasing. There is a brief concluding section. An Appendix provides selected derivations and formulas. The complete working paper (Seidman 1983*a*), available from the author, provides all derivations, so that the text can focus on assumptions and results.

I. The Steady State of a Life-Cycle Growth Model with a Bequest Motive and Taxes

I have analyzed the steady state elsewhere (Seidman 1983*b*). This section, therefore, will present only the results necessary to understand the transition path analysis of this paper.

For tractability, it is assumed a bequest is given only at death and an inheritance is received only at labor force entry. All ages are "work ages," expressed relative to the labor force entry age of zero. Every person is identical except for age; the impact of this assumption on the analysis of conversion is given elsewhere (Seidman and Maurer 1982). The representative person, on entry, chooses a (C_t, K_J) path to maximize his lifetime utility function (1) subject to his expected lifetime budget constraint (2) (see Strotz [1956] and Atkinson [1971] for discussions of the lifetime utility function):

$$U = \sum_{t=1}^{J} \frac{\ln C_t}{(1 + \rho)^t} + \beta \frac{\ln K_J}{(1 + \rho)^J}, \tag{1}$$

$$\sum_{t=1}^{J} \frac{C_t/(1 - t_c)}{(1 + r_n)^t} + \frac{K_J}{(1 + r_n)^J} = H + \sum_{t=1}^{R} \frac{w_{nt}}{(1 + r_n)^t}, \tag{2}$$

where $r_n \equiv (1 - t_r)r$, $w_{nt} \equiv (1 - t_w)w_t$, U = lifetime utility, C_t = consumption at age t, K_J = bequest at death, ρ = subjective discount rate, β = taste for bequest parameter, J = age at death (last age of life), r = rate of interest, H = inheritance at labor force entry, w_t = wage at age t, R = age of retirement (last age of work), t_c = consumption tax rate (ratio of tax to the sum of consumption plus tax; hence, $C_t/[1 - t_c]$ is the sum of consumption plus tax), t_w = labor income tax rate, and t_r = capital income tax rate. For a Y tax, $t_w = t_r = t_y$, and t_c = 0; for a C tax, $t_r = t_w = 0$.

In this study, it is assumed that a bequest is *not* treated as taxable consumption under a C tax. This assumption is incorporated in (2), where the bequest, K_J, is not subject to tax. Whether bequests are exempt or taxable under a C tax has a direct bearing on the pattern of transitional losses; a bequest motive will reduce the transitional burden of older cohorts only if a bequest is tax exempt under the C tax. (See Seidman [1980] for another rationale for such an exemption.)

For the issues investigated in this paper, the logarithmic utility function, (1), is probably sufficiently flexible. Both Summers (1979) and Auerbach-Kotlikoff (1983) use the logarithmic utility function for their transition path analysis. By varying ρ, the subjective discount rate, (1) permits the preference for present consumption relative to future consumption to vary (raising ρ raises the relative preference for present consumption). Nevertheless, it is possible that the conclusions may prove sensitive to this functional form.

β is the bequest motive parameter. There are two ways to understand the implications of the magnitude of β. It can be shown that β is the ratio of a person's bequest to his last-age consumption-plus-consumption tax. For example, a $\beta = 10$ implies that the person leaves a bequest 10 times the size of his consumption-plus-consumption tax at age J. Second, it can be shown that the share of a person's lifetime wealth allocated to the bequest depends on β, ρ, and J. Consider two examples that will be used in Sections IV and V. Given $J = 55$, $\rho = .08$ and $\beta = 10$ imply that 1.1 percent of lifetime wealth is allocated to the bequest; $\rho = .08$ and $\beta = 20$ imply that 2.3 percent is allocated.

The constrained maximization implies that each person will choose a consumption path with a growth rate of $(r_n - \rho)/(1 + \rho)$. Two special cases should be noted. If $r_n = \rho$, then C_t will be constant over the person's life. If $\rho = 0$, then C_t will grow at the rate r_n.

Although each person takes his inheritance as given when choosing (C_t, K_J), in the steady state of the growth model, total inheritances received by the entering cohort must equal total bequests given by the dying cohort.

Government expenditure is financed by equal tax revenue and is assumed to consist solely of government consumption so that the capital stock results solely from individual saving. There is no role for debt policy.

Production is Cobb-Douglas, and each factor is pa[illegible] product:

$$q = k^{\alpha}, \quad 0 < \alpha < 1,$$

where q = output per effective labor (Q/E) and k = [illegible]tive labor (K/E);

$$r = \alpha k^{\alpha-1},$$

where r = the interest rate;

$$w_e = (1 - \alpha)k^{\alpha},$$

where w_e = wage per effective labor.

There is labor-augmenting technical change at the constant exogenous rate g, so that effective labor per worker grows at g; and the labor force grows at the constant exogenous rate n, so that total effective labor, capital, and output grow at the rate $(n + g + ng)$.[1] Depreciation is assumed to be zero, so that aggregate saving equals the increase in the capital stock, and the saving to capital ratio also equals ($n + g + ng$). At any time v, each worker, whatever his age, receives the same wage—equal to the marginal product of a worker in the economy at time v. Capital per effective labor (k), the wage per effective labor (w_e), and the interest rate (r) are constant. The wage per worker grows at the rate g, both over time and over each person's life.

---

[1] Note that with discrete period growth rates n and g the growth rate of effective labor, G, is given by $(1 + G) = (1 + n)(1 + g)$.

By aggregating the saving of each individual at each age and using the relationships (1)–(5) and the constraints previously described, it can be shown that the steady-state r_n of the economy is given implicitly by (6), which is kept relatively concise by a summation function M:

$$\frac{r_n}{(1 - t_r)} = \frac{\left(\frac{\alpha}{1 - \alpha}\right)(r_n - n - g - ng)}{(1 - t_w)\left\{\frac{M_1^R[(1 + g)/(1 + r_n)] \cdot M_1^J(y_n)}{D \cdot M_1^R[1/(1 + n)]} - 1\right\}},$$

$$M_L^{L'}(X) \equiv \sum_{j=L}^{L'} X^j = \frac{X^{L'+1} - X^L}{X - 1}, \quad X \neq 1, \tag{6}$$

$$D \equiv M_1^J\left(\frac{1}{1 + \rho}\right) + \beta\left[\left(\frac{1}{1 + \rho}\right)^J - (y_n)^J\right],$$

$$y_n \equiv \frac{(1 + r_n)}{(1 + g)(1 + n)(1 + \rho)},$$

where α = the Cobb-Douglas exponent (in eq. [3]), n = growth rate of the labor force, and g = growth rate of effective labor per worker (rate of exogenous labor-augmenting technical change).

Given a set of parameters and tax rate values, a computer-implemented iterative procedure is used to obtain r_n. Then k is given by (4); note that r and k always vary inversely.

It can be shown that the saving rate, b, is given by

$$b = (n + g + ng)k^{1-\alpha}, \tag{7}$$

$$b \equiv \frac{s}{q}, \tag{8}$$

where s = saving per effective labor (S/E). Note that b varies directly with k; the elasticity is $(1 - \alpha)$.

In (6), t_c does not appear. In response to a C tax, each person finds it optimal to reduce his C_t such that the sum of consumption-plus-consumption tax, $C_t/(1 - t_c)$, stays constant, so that saving at each age, S_t, is unaffected. Hence, the steady-state k and r are the same as under no tax (recall that we have assumed there is no government saving so the capital stock results solely from individual saving).

It is also clear from (6) that a Y tax ($t_w = t_r = t_y$) has no effect on r_n. Hence, a Y tax unambiguously raises gross r and reduces k. Thus, in this model (with its functional forms) a C tax must achieve a higher k than a Y tax.

In fact, if the Y tax rate is initially t_y (e.g., 30 percent), conversion to a C tax would reduce r by t_y (30 percent), for any value of t_c, for the following reason. From (6), it is clear that, beginning at $t_r = t_w = t_y$ (30 percent), reducing these rates to 0 percent would have no effect on r_n, which is t_y (30 percent) less than the initial r. Since t_c does not appear in (6), imposing a C tax would not affect r_n, which would equal r under a C tax because $t_r = 0$ percent. Hence, r under a C tax would be t_y (30 percent) less than under a Y tax. Note that this result holds for any ρ, β, and t_c.

It can be shown that if the Y tax rate is initially t_y, conversion to a C tax with any t_c would have the following impact on k, s, b, and q:

$$\frac{\Delta k}{k} = \frac{\Delta s}{s} = \left(\frac{1}{1 - t_y}\right)^{1/(1-\alpha)} - 1, \tag{9}$$

$$\frac{\Delta b}{b} = \left(\frac{1}{1 - t_y}\right) - 1, \tag{10}$$

$$\frac{\Delta q}{q} = \left(\frac{1}{1 - t_y}\right)^{\alpha/(1-\alpha)} - 1. \tag{11}$$

For $t_y = 30$ percent, $\Delta k/k = \Delta s/s = 66$ percent, $\Delta b/b = 43$ percent, and $\Delta q/q = 19$ percent.

Earlier it was noted that the maximization of (1) subject to
plies that each person will choose a lifetime consump
the growth rate $(r_n - \rho)/(1 + \rho)$. Since both a Y tax and
t_y and t_c, achieve the same r_n, the growth rate will be t
both taxes. This fact will greatly simplify welfare comp
cially in the no-bequest case.

While (9)–(11) imply that the percentage changes in k
independent of ρ, β, and t_c, this is not true for consump
effective labor, c, which obtains a maximum at the "golden rule k
Although the C tax always achieves a higher k than the Y tax, the "golden rule" literature makes clear that a higher k does not necessarily imply a higher c, or higher lifetime welfare.

Nevertheless, for all parameters tried yielding an r greater than the natural growth rate of the economy $(n + g + ng)$, a C tax did achieve a higher lifetime utility for the representative person than the Y tax, when tax rates were set to achieve the same revenue per effective labor. Because the purpose of this study is to assess possible transitional harm on the way to steady-state gain, and also because it is probably empirically realistic to assume that $r > n + g + ng$, all simulations are based on parameter sets for which conversion to a C tax would raise steady-state lifetime utility.

The change in steady-state lifetime welfare when the Y tax is converted to a C tax is meaningfully measured by δ_L, the percentage change in the Y tax lifetime consumption-bequest vector ($J + 1$ elements) that would be required to achieve the same lifetime utility that occurs under the C tax.

It can be shown that in the no-bequest case, δ_L is simply equal to the percentage change in aggregate consumption per effective labor, c. This simple relationship is due to the fact that the growth rate of consumption over a person's life is the same under both taxes because r_n is the same. If only consumption affects utility (no bequest), then $(1 + \delta_L)$ is the ratio of C_{ct} (consumption at age t under the C tax) to C_{yt} (consumption at age t under the Y tax) at every age t over the person's life, and aggregation of consumption across cohorts at any time v preserves the $(1 + \delta_L)$ ratio.

All results presented in this paper use the following parameter values: $n = 1$ percent, $g = 2$ percent ($n + g + ng = 3.02$ percent); $\alpha = .3, R = 45, J = 55$. Under the Y tax, $t_y = 30$ percent; t_c is then set to achieve the same revenue per effective labor.

II. The Simulation Model

Once the economy leaves the steady state, it is no longer true, when a person replans the remainder of his life at time v and age t, that he will continue along the path he planned at entry. To select a rest-of-life path at age t, time v, the person must form expectations concerning the future paths of three variables: $w(j)$, $r(j)$, and tax rates $t_i(j)$, where j indicates his age.

Auerbach and Kotlikoff (1983) assume "perfect foresight," where expectations along the transition path prove correct. Summers (1979) assumes "steady-state expectations" that would prove correct if and only if the economy were in a steady state; each person is assumed to act as if he expects the interest rate to remain constant at its current value, the wage per worker to grow at the rate g, and tax rates to remain at their current values. Comparison of the simulation results of these two studies suggests that the conclusions of this paper are unlikely to depend on which expectations hypothesis is utilized, because the basic qualitative pattern is similar. This paper adopts "steady-state" expectations because, to me, it seems more realistic (as well as simpler to implement).

Each period, then, a person of age t plans C_j for $j = t, \ldots, J$ and K_J by maximizing rest-of-life utility subject to lifetime wealth at age t. Although each person plans the remainder of his life each period, only the initial $C_t(v)$ is relevant for the simulation model, because each person will replan again next period.

For a person age t at time v, his income is computed as the sum of his wage and capital income. Subtracting his tax and consumption yields his saving. Summing over the population yields total saving, which constitutes the increase in the capital stock. Inserting the new capital and effective labor into the Cobb-Douglas production function, computing marginal products, yields the new interest rate and wage. For each person, adding saving to wealth yields new wealth (the bequests of the dying cohort become the inherited wealth of the cohort just entering the work force). The process is repeated each period, generating the path of the economy.

Each simulation begins from an initial steady state, with the wealth held by a person of each age consistent with that steady state. When the simulation model was run for the same parameter and tax values that determined that initial steady state, the economy remained in that steady state in each subsequent period. When the tax rates were changed and then a simulation was run, in every case tried the economy followed a transition path that eventually converged to the new steady state associated with the new tax rates as given by (6). Thus, the steady state appears to be stable.

Each simulation begins from an initial Y tax steady state. The C tax is introduced, and the tax rate, $t_c(v)$, is adjusted every period v to keep tax revenue per effective labor equal to its constant value under the Y tax steady state. The economy gradually converges to the higher k given by (6) and (4).

For each age cohort working or retired when convers
well as each that enters after conversion, this study
change in "rest-of-life" welfare, measured by δ_{Rt}, th
change in the Y tax "rest-of-life" consumption-bequest
$-$ t elements for a person of age t when conversion occ
be required to achieve the same "rest-of-life" utility it a
conversion to the C tax. It should be noted that the youn
cohort, the less the discrepancy between δ_{Rt} and δ_l (the percentage change in *lifetime* welfare); for a cohort just entering the work force at the time of conversion, as well as all cohorts who will enter in the future, $\delta_{Rt} = \delta_L$. As the economy approaches its new steady state the $\delta_{Rt} = \delta_L$ of future cohorts will converge to the steady-state δ_l defined in Section I.

III. Conversion in a No-Bequest Economy

To isolate the impact of a bequest motive on conversion, we begin with the no-bequest case. Three values of ρ are considered. Under a Y tax (t_y = 30 percent), (6) yields the corresponding steady-state r for

TABLE 1

CONVERSION IN A NO-BEQUEST ECONOMY

	Y Tax Steady State	First Year of Conversion	%Δ upon Conversion	C Tax Steady State	%Δ between Steady States
$\rho = .0001$:					
r	6.3%	6.3%	0	4.4%	−30
k	9.28	9.28	0	15.45	+66
b	14%	39%	+173	21%	+43
q	1.95	1.95	0	2.27	+17
T	.59	59	0	.59	0
s	.28	76	+173	.46	+66
c	1.08	.60	−44	1.22	+13
δ_L	. . .	. . .	. . .	.	+13
$\rho = .052$:					
r	12.1%	12.1%	0	8.5%	−30
k	3.67	3.67	0	6.10	+66
b	8%	33%	+329	11%	+43
q	1.48	1.48	0	1.72	+17
T	.44	.44	0	.44	0
s	11	.49	+329	18	+66
c	93	.55	−41	1.10	+18
δ_L				. .	+18
$\rho = .110$:					
r	20.2%	20.2%	0	14.2%	−30
k	1.76	1.76	0	2.92	+66
b	4%	30%	+635	6%	+43
q	1.18	1.18	0	1.38	+17
T	.36	.36	0	36	0
s	.05	36	+635	09	+66
c	77	.46	−40	93	+21
δ_L			.	. . .	+21

each ρ. As indicated in table 1, for a "low" ρ (.0001), $r = 6.3$ percent; for a "moderate" ρ (.052), $r = 12.1$ percent; for a "high" ρ (.110), $r = 20.2$ percent. As might be expected, a higher ρ, indicating a greater relative preference for present over future consumption, results in a lower capital intensity k and saving rate b and a higher interest rate (marginal product of capital) r.

Before examining the transition path, it is useful to focus on the final steady state. Table 1 presents these results in the right-hand column. As shown in Section I, the percentage changes in r, k, b, q, and s are independent of ρ and were given in (9)–(11). As also explained in Section I, c and δ_L do depend on ρ, and the percentage change in c is equal to δ_L. As table 1 indicates, r is always greater than $(n + g + ng) = 3.02$ percent, so that the economy is always at a $k < k_g$ (even with $\rho = .0001$ under the C tax, when r is 4.4 percent); it is not surprising that the percentage change in c, which equals δ_L, is positive for all three ρ. Moreover, the higher is ρ, and therefore the greater is the gap between the initial k_y and k_g, the greater is the percentage increase in c and percentage increase in welfare, δ_L. For a "moderate" ρ (.052) with an initial r = 12.1 percent, the percentage increase in c is 18 percent.

Of course, it is not possible to move instantly from the Y tax steady state to the C tax steady state. A central purpose of this paper is to focus on the short-run reduction in consumption per effective labor that must follow conversion and to assess the impact on the welfare of particular age cohorts along the transition path.

The second and third columns of table 1 give the change in the economy in the year of conversion. With the C tax rate set so that tax revenue per effective labor remains at its value under the Y tax steady state, the utility function implies that the population will choose to reduce its consumption drastically; for all three ρ, consumption per effective labor is reduced just over 40 percent. It then climbs steadily until it converges to its steady-state value. As will be indicated in Section IV, this 40 percent reduction persists when a bequest motive is introduced.

It should be noted that a similar drastic reduction in consumption also occurs in both the Summers and Auerbach-Kotlikoff studies. To isolate the importance of a bequest motive and age phasing, this study follows the two previous studies closely, including the choice of utility function. Since many would regard so drastic a decline in consumption as implausible, it would be important for future rese[illegible] periment with alternative forms of the utility function

The higher is ρ, and therefore r, the faster the econo[illegible] to its new steady state. In all cases, the economy mildly [illegible] steady state and exhibits damped oscillations as it conv[illegible]

For a "low" ρ (.0001) and "low" initial r (6.3 perce[illegible] percent in the year of conversion; not until year 19 [illegible] exceed its initial value. The new steady-state r is first reached [illegible] 30; half the distance is covered by year 5. For a "moderate" ρ (.052) and "moderate" initial r (12.1 percent), c is cut 41 percent in the year of conversion; the initial value for c is surpassed in year 8. The new steady-state r is first reached in year 14; half the distance is covered by year 3. For a "high" ρ (.110) and "high" initial r (20.2 percent), c is cut 40 percent in the year of conversion; the initial value for c is surpassed in year 4. The new steady-state r is first reached in year 8; half the distance is covered by year 2.

Figure 1 shows the change in rest-of-life welfare, δ_R, for each ρ. The higher is ρ, the more favorable (less unfavorable) is the impact on all cohorts. Yet even for the highest ρ, the transitional losses are significant. Everyone older than work age 27 at the time of conver-

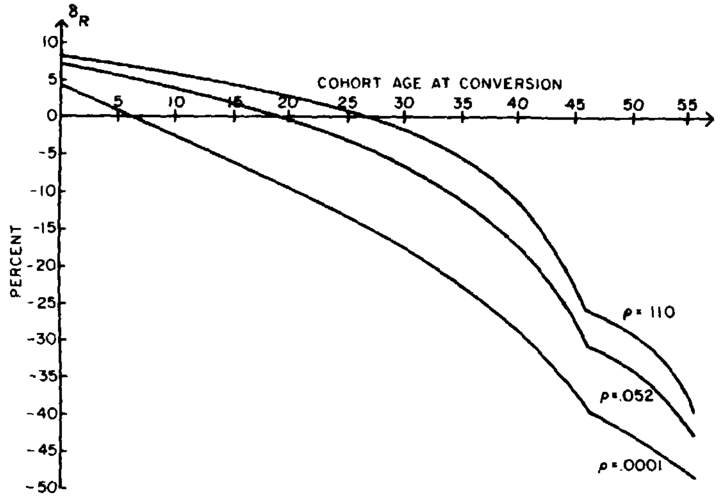


FIG. 1.—The change in cohort "rest-of-life welfare" due to abrupt conversion in a no-bequest economy with alternative values of ρ.

sion—more than half of all cohorts—has rest-of-life welfare reduced. The maximum loss, incurred by the oldest cohort, is 40 percent of its rest-of-life (last year) welfare. Hence, the average loss of cohorts over 27 is roughly 20 percent of rest-of-life welfare.

Why are the losses for older cohorts so large in the no-bequest economy? The growth rate of consumption over a person's life is $(r_n - \rho)/(1 + \rho)$: 4.4 percent for $\rho = .0001$, 3.1 percent for $\rho = .052$, and 2.8 percent for $\rho = .110$. This implies that the ratio of last age ($t = J = 55$) to first age ($t = 1$) consumption is 10.25 for $\rho = .0001$, 5.19 for $\rho = .052$, and 4.55 for $\rho = .110$. Hence, conversion to a C tax imposes a relatively large tax burden on high-consuming older cohorts.

What about the young? Even in the high ρ case, the δ_R (8.0 percent) of the youngest cohort, which just enters at the time of conversion, is still well below the steady-state δ_L (21 percent). Why is this so? This youngest cohort begins its work career with w_e at the Y tax value; 8 years must pass before w_e first reaches the higher C tax steady-state value. Hence, for nearly one-fifth of its work life, this cohort earns a w_e less than the C tax value. On the other hand, it also receives an r_n greater than the C tax steady-state value for these same 8 years. Evidently, the former outweighs the latter.

IV. Conversion in an Economy with a Bequest Motive

This section examines conversion in a "moderate" r economy, where the initial Y tax steady-state r of 12 percent is generated by alternative pairs, (β, ρ). With an initial Y tax rate of 30 percent, the following four pairs each generated an initial steady-state r of 12 percent: (.0001, .052), (5.0, .067), (10.0, .077), (20.0, .087).

As the (β, ρ) pairs indicate, β and ρ must both be raised to hold r constant at 12 percent. This is because raising β, the bequest motive (holding all other parameters constant), always raised steady-state k and b, as might be expected, thereby reducing r. On the other hand, raising ρ, the relative preference for present over future consumption (holding all other parameters constant), always reduced steady-state k and b, as might be expected, thereby raising r.

Thus, if the economy is observed to be in an initial Y tax steady state (t_y = 30 percent) with r = 12 percent, no information is conveyed concerning the strength of the bequest motive. It will be shown, however, that when conversion to a C tax occurs, the impact on the welfare of different age cohorts will depend on the particular (β, ρ) that generated the initial Y tax steady state.

Before examining the transition path, it is useful to focus on the final steady state. As shown in Section I, the percentage changes in r, k, b, q, and s are independent of (β, ρ). Since r is 12 percent in all four cases under the 30 percent Y tax, $\%\Delta c$ is always approximately 18 percent, as shown in table 1, for the r = 12.1 percent case. Although δ_l does not exactly equal $\%\Delta c$ because the bequest affects utility, in fact it does so approximately. For all four (β, ρ), δ_l is approximately the same as $\%\Delta c$: 18–19 percent. Hence, δ_l is insensitive to the particular (β, ρ) combination that generates the same initial r = 12 percent.

For all (β, ρ) combinations, the reduction in c in the year of conversion was within a few percentage points of 40 percent, just as in the no-bequest economy. Given this same aggregate cutback in consumption, the important question becomes, Does a bequest m[illegible] a reallocation of the consumption sacrifice among coho[illegible] maximum loss suffered by any cohort is reduced?

Figure 2 shows that the higher are the β and ρ generat[illegible] r = 12 percent, the smaller is the maximum loss su[illegible] cohort. Each curve shows the change in rest-of-life we[illegible] each age cohort for a particular (β, ρ). The (.0001, .052) [illegible] already plotted in figure 1. But now, the higher are (β, ρ), the less negative are the δ_R; in fact, in the β = 20.0 case, the oldest cohort actually gains from conversion.

Why does a higher (β, ρ) reduce the maximum loss? Recall that, in the Y tax steady state, the growth rate of consumption over the person's life is $(r_n - \rho)/(1 + \rho)$. Hence, the higher are β and ρ, given the initial r_n = 8.4 percent, the lower is the growth rate and the flatter is the consumption path. Table 2 shows the growth rate under each (β, ρ) and the resulting ratio of last age ($t = J = 55$) to first age ($t = 1$) consumption under each path. For β = 20.0, ρ = .087, the growth

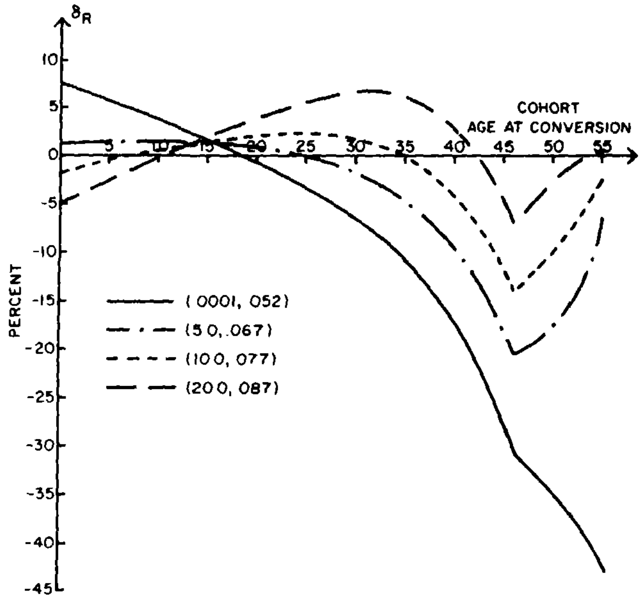


FIG. 2.—The change in cohort "rest-of-life welfare" due to abrupt conversion in alternative (β, ρ) economies yielding an initial Y tax r of 12 percent

rate is slightly negative, and the ratio is less than 1 (.88). The flatter the consumption path, the smaller will be the tax burden imposed on older cohorts immediately following conversion.

As figure 2 shows, however, a flatter path, while helping older cohorts under conversion, hurts younger cohorts. In the β = .0001 case, older cohorts are hurt severely, while younger cohorts gain. For β = 20.0, while the losses to older cohorts are greatly reduced, the youngest cohorts now lose. A flatter path reallocates the burden of conversion away from old to young cohorts. This should not be sur-

TABLE 2

THE STEADY-STATE LIFETIME CONSUMPTION PATH IN AN ECONOMY WITH A BEQUEST MOTIVE

β	ρ	r_n (%)	Growth Rate of Consumption (%) $(r_n - \rho)/(1 + \rho)$	Ratio of Last Age to First Age Consumption $C_{55}/C_1 = [(1 + r_n)/(1 + \rho)]^{54}$
.0001	.052	8.456	3.1	5.19
5.0	.067	8 384	1.6	2.33
10.0	0765	8.460	8	1.50
20.0	.087	8.434	−.2	.88

prising because, for all (β, ρ), the first-year reduction in c is approximately 40 percent, so that a smaller loss for some entails a greater loss for others. This reallocation, however, does significantly reduce the maximum loss in rest-of-life welfare suffered by any cohort.

Convergence is also somewhat slower in the high-β, high-ρ economy. While two-thirds of the gap between $r = 12$ percent and $r = 8.4$ percent is closed in 5 years for all four values of (β, ρ), for $\beta = 20.0$ the new steady state is not reached until year 29; for $\beta = .0001$, it is reached in year 14. Slower convergence is another reason why younger cohorts are hurt. Once the new steady state is reached, cohorts that enter thereafter will enjoy the steady-state δ_L. But for high (β, ρ), cohorts that enter several years after conversion will be hurt.

V. Conversion with Age Phasing

This section examines conversion in a "moderate" r economy (12 percent under the 30 percent Y tax), with a "moderate" $(\beta, \rho) = (10.0, .077)$ under a policy of age phasing. Recall that the steady-state δ_L is 19 percent.

Under the method of age phasing examined here for illustration, it is assumed that the population is divided at age D. Persons older than D remain subject to the Y tax for the remainder of their lives. Persons with age less than or equal to D are immediately and permanently subject to the C tax. Each year, the dividing age is increased by one year, until the whole population is subject to the C tax.

There is one policy parameter to be chosen for each si
dividing age D. The t_y that will apply to the older, pr
for the remainder of their lives is kept at its initial Y
percent). Given this t_y, each year the $t_c(v)$ for the young
by a computer-implemented iterative procedure so that
nue per effective labor is the same as it would have be
steady state.

In contrast to the bequest motive, age phasing does reduce percentage reduction in c that occurs in the year of conversion. The cohorts older than D remain under the Y tax, at the same $t_y = 30$ percent; they do not suffer a tax increase, nor are they given an incentive to reduce their consumption. As D declines, the share of the population so "protected" rises. For $D = 55$ (no age phasing), the $\%\Delta c$ in the year of conversion would be -41 percent; for $D = 45$, -37 percent; for $D = 35$, -30 percent; and for $D = 25$, -23 percent.

Figure 3 shows that age phasing reduces the maximum loss suffered by any cohort. Four values of D are considered: 55 (in effect, no age phasing), 45, 35, and 25. The curve $D = 55$ is also plotted in figure 2; the maximum loss, δ_R, is 14 percent, suffered by the cohort

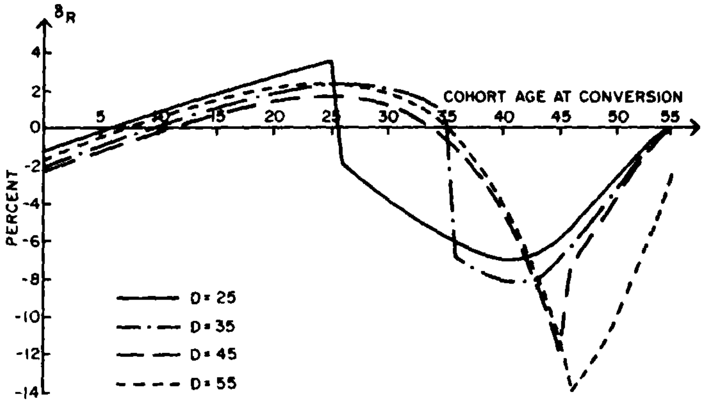


FIG. 3.—The change in "rest-of-life welfare" due to conversion with alternative age phasing (D) in a "moderate" (β, ρ) economy.

with age 46. As D is reduced to 25, the maximum loss declines to 7 percent, and the age of the cohort experiencing that loss also declines to 41.

The reduction in the maximum loss is not achieved costlessly, however. The number of cohorts that lose rises as D declines. Also, age phasing reduces the speed of convergence. With $D = 55$, r closes two-thirds of the gap between 12 percent and 8.4 percent in 5 years; with $D = 25$, in 10 years. Thus, with age phasing, a longer time will elapse before a cohort enters that will enjoy the steady-state gain, $\delta_L = 19$ percent.

VI. Conclusions

This paper analyzes conversion from an income tax to a consumption tax (which exempts bequests) in a life-cycle growth model. In contrast to two previous studies by Summers and Auerbach-Kotlikoff, this study includes a bequest motive and considers a policy of age phasing. In this particular model, with its production and utility functions, parameter values, and "steady-state expectations," the following main results were obtained:

1. In the absence of a bequest motive or a policy of age phasing, conversion to a C tax imposed significant losses of "rest-of-life welfare" on older cohorts, even when a C tax would significantly raise steady-state lifetime welfare. Exclusive focus on steady states is therefore inadequate and misleading when capital accumulation is influenced by a policy change.

2. Alternative combinations of bequest motive and subjective discount rate can generate the same steady-state capital intensity under

the initial income tax. For example, raising both by the proper magnitude can hold capital intensity constant, because a greater bequest motive would in itself raise capital intensity, while a greater discount rate would in itself lower capital intensity.

The greater the bequest motive and discount rate generating a given Y tax capital intensity, the smaller will be the transitional losses imposed on older cohorts during conversion to a C tax, provided the C tax exempts bequests. Moreover, over the relevant range, a greater bequest motive discount-rate combination also reduced the maximum loss suffered by any cohort. Thus, a life-cycle growth model without bequests will overstate the maximum cohort loss that results from conversion. Because the result depends crucially on the exemption of bequests under the C tax, it may be interpreted as providing one argument for such exemption.

3. A policy of age phasing can reduce the transitional losses imposed on older cohorts, and the maximum loss suffered by any cohort, under conversion.

4. In this particular life-cycle growth model, even with a bequest motive and age phasing, transitional losses remain. Yet, like the Summers and Auerbach-Kotlikoff models, this one contains several features that tend to exaggerate transitional harm—such as the assumption that a person's consumption grows as rapidly in retirement as during work years. Until models that modify these convenient but implausible features are developed, it would be premature to draw firm conclusions concerning the magnitude of the trans
that would result from conversion. The main task of
been to demonstrate that incorporating two featur
nored—a bequest motive and age phasing—can sign
transitional losses. Further research is required to dete
the transitional problem is as serious as has been
gested.

Appendix

This Appendix gives selected derivations and formulas. The complete working paper (Seidman 1983*a*), available from the author, gives all derivations

Steady State

For each person, (C_t, K_J) is given by:

$$\frac{C_t}{1 - t_t} = \left(\frac{1}{\lambda}\right)\left(\frac{1 + r_n}{1 + \rho}\right)^t, \tag{A1}$$

$$K_J = \left(\frac{1}{\lambda}\right)\beta\left(\frac{1 + r_n}{1 + \rho}\right)^J, \tag{A2}$$

where λ = the Lagrange multiplier.

Hence, C_t grows at the rate $(r_n - \rho)/(1 + \rho)$, and β is the ratio of the bequest to last-age consumption-plus-consumption tax.

$$\frac{K_J/(1 + r_n)^J}{I} = \frac{1}{1 + \left[M_1^J\left(\frac{1}{1 + \rho}\right) \Big/ \beta\left(\frac{1}{1 + \rho}\right)^J\right]}, \tag{A3}$$

where $I = H + \Sigma_{t=1}^{R}\,[w_{xt}/(1 + r_n)^t]$ = lifetime wealth and M is defined with (6). Equation (A3) gives the fraction of lifetime wealth devoted to a bequest.

At time v total bequests equals total inheritances:

$$P_J(v) \cdot K_J(v) = P_0(v) \cdot H(v), \tag{A4}$$

where $P_t(v)$ = population of age t at time v. This implies that for each person

$$H = [(1 + g)(1 + n)]^{-J}K_J. \tag{A5}$$

Using (A5),

$$\frac{1}{\lambda} = \frac{w_{x0}M_1^R[(1 + g)/(1 + r_n)]}{D}, \tag{A6}$$

where w_{x0} = net wage in the economy when a person was work age 0 and D is defined with (6).

Aggregating over the whole population at time v:

$$\frac{C(v)/(1 - t_c)}{w_x(v)L(v)} = \frac{M_1^R[(1 + g)/(1 + r_n)] \cdot M_1^J(y_n)}{D \cdot M_1^R[1/(1 + n)]}, \tag{A7}$$

where $C(v)$ = aggregate consumption at time v, $w_x(v)L(v)$ = aggregate net labor income at time v, and y_n is defined with (6). In the steady state,

$$w_xL + r_nK - \frac{C}{1 - t_c} \equiv S \equiv (n + g + ng)K. \tag{A8}$$

Using (A7), (A8), (4), and (5) yields (6). Individual wealth is given by

$$A(t + 1) = (1 + r_n)^t\left[H + \sum_{j=1}^{x} \frac{(1 - t_w)w_j}{(1 + r_n)^j} - \sum_{j=1}^{t} \frac{C_j/(1 - t_c)}{(1 + r_n)^j}\right], \tag{A9}$$

where $A(t + 1)$ = wealth at the beginning of age $t + 1$, $x = t$ for $1 \leq t \leq R$, $x = R$ for $R + 1 \leq t \leq J$, and $A(1) = H$. At the beginning of time $(v + 1)$,

$$\begin{aligned} A_{v+1}(t + 1) = {} & \frac{(1 + r_n)^t w_{xt}(1 + g)^{v-t}}{D} \\ & \times \left\{[(1 + g)(1 + n)]^{-J} M_1^R\left(\frac{1 + g}{1 + r_n}\right) \cdot \beta\left(\frac{1 + r_n}{1 + \rho}\right)^J \right. \\ & \left. + DM_1^x\left(\frac{1 + g}{1 + r_n}\right) - M_1^t\left(\frac{1}{1 + \rho}\right) \cdot M_1^R\left(\frac{1 + g}{1 + r_n}\right)\right\}, \end{aligned} \tag{A10}$$

where $A_{v+1}(1) = H$;

$$T_c = t_c[k_c^\alpha - (n + g + ng)k_c], \tag{A11}$$

$$T_\gamma = t_\gamma k_\gamma^\alpha, \tag{A12}$$

where T_i = tax revenue per effective labor under tax i;

$$U_{\bar{v}} = \sum_{t=1}^{J} \frac{\ln\left(\dfrac{\{(1 - t_c)w_{xe}(1 + g)^{\bar{v}}M_1^R[(1 + g)/(1 + r_n)]\}[(1 + r_n)/(1 + \rho)]^t}{D}\right)}{(1 + \rho)^t}$$

$$+ \frac{\beta \ln\left\{\dfrac{w_{xe}(1 + g)^{\bar{v}}M_1^R[(1 + g)/(1 + r_n)]\beta[(1 + r_n)/(1 + \rho)]^J}{D}\right\}}{(1 + \rho)^J}, \qquad \text{(A13)}$$

where $U_{\bar{v}}$ = lifetime utility of a person who enters at time $\bar{v}$ δ_t is defined by:

$$\sum_{t=1}^{J} \frac{\ln[(1 + \delta)C_{ty}]}{(1 + \rho)^t} + \frac{\beta \ln\{(1 + \delta)K_{Jy}\}}{(1 + \rho)^J} = U_{\bar{v}c}, \qquad \text{(A14)}$$

where $U_{\bar{v}c}$ = lifetime utility under the C tax and (C_{ty}, K_{Jy}) = the path chosen under the Y tax. Using the fact that r_n is the same under both the Y tax and the C tax:

$$\delta = \exp\left\{\frac{M_1^J\left(\dfrac{1}{1 + \rho}\right) \cdot \ln\left[\dfrac{(1 - t_c)w_{ve}}{(1 - t_y)w_{ye}}\right] + \beta\left(\dfrac{1}{1 + \rho}\right)^J \cdot \ln\left[\dfrac{w_{ve}}{(1 - t_y)w_{ye}}\right]}{M_1^J\left(\dfrac{1}{1 + \rho}\right) + \beta\left(\dfrac{1}{1 + \rho}\right)^J}\right\} - 1 \qquad \text{(A15)}$$

From (A14) and (A7), with $\beta = 0$,

$$1 + \delta_t = \frac{C_c(v)}{C_y(v)}.$$

Hence, in the no-bequest case δ_t equals the percentage ch
consumption per effective labor.

The Simulation Model

At time v a person of age t plans C_j for $j = t, \ldots, J$ and K_J to maximize (1-s) subject to (2-s):

$$U = \sum_{j=t}^{J} \frac{\ln C_j}{(1 + \rho)^{j-t}} + \beta \frac{\ln K_J}{(1 + \rho)^{J-t}}, \qquad \text{(1-s)}$$

$$\sum_{j=t}^{J} \frac{C_j/[1 - t_c(v)]}{[1 + r_n(v)]^{j-t}} + \frac{K_J}{[1 + r_n(v)]^{J-t}}$$

$$= A_v(t)[1 + r_n(v)] + Z \sum_{j=t}^{R} \frac{[1 - t_w(v)]w(v)(1 + g)^{j-t}}{[1 + r_n(v)]^{j-t}}. \qquad \text{(2-s)}$$

where $Z = 1$ for $1 \leq t \leq R$ and $Z = 0$ for $R + 1 \leq t \leq J$. Hence:

$$\frac{C_t(v)}{1 - t_c(v)} = \frac{1}{\lambda^*} = \frac{A_t(t)[1 + r_n(v)] + Z\left\{[1 - t_w(v)]w(v)\left[\frac{1 + r_n(v)}{1 + g}\right]^t M_t^R\left[\frac{1 + g}{1 + r_n(v)}\right]\right\}}{(1 + \rho)^t M_t^J\left(\frac{1}{1 + \rho}\right) + \beta\left(\frac{1}{1 + \rho}\right)^{J - t}}, \tag{A17}$$

where $Z = 1$ for $1 \leq t \leq R$ and $Z = 0$ for $R + 1 \leq t \leq J$;

$$K_J(v) = \frac{\beta}{\lambda^*} = \frac{\beta A_v(J)[1 + r_n(v)]}{(1 + \beta)}. \tag{A18}$$

Using (A4),

$$H(v) = \frac{\beta A_v(J)[1 + r_n(v)]}{(1 + \beta)(1 + n)^J}. \tag{A19}$$

δ_t is defined by

$$\sum_{j=t}^{J} \frac{\ln (1 + \delta_t)C_{jy}}{(1 + \rho)^{j - t}} + \frac{\beta \ln (1 + \delta_t)K_{Jy}}{(1 + \rho)^{J - t}} = U''_{t\bar{v}c}, \tag{A14-s}$$

where $U''_{t\bar{v}c}$ = "rest-of-life" utility of a person of age t at time $\bar{v}$ under the C tax and (C_{jy}, K_{Jy}) = the path chosen under the Y tax. Then:

$$\delta_t = \left(\exp\left\{\frac{U''_{t\bar{v}c} - U''_{t\bar{v}y}}{(1 + \rho)^t\left[M_t^J\left(\frac{1}{1 + \rho}\right) + \beta\left(\frac{1}{1 + \rho}\right)^J\right]}\right\}\right) - 1.$$

$(U''_{t\bar{v}c} - U''_{t\bar{v}y})$ turns out to be independent of $\bar{v}$, so that δ_t does not depend on the time $\bar{v}$ chosen to begin conversion.

References

Atkinson, Anthony B. "Capital Taxes, the Redistribution of Wealth and Individual Savings." *Rev. Econ. Studies* 38 (April 1971): 209–27.

Auerbach, Alan J., and Kotlikoff, Laurence J. "National Savings, Economic Welfare, and the Structure of Taxation." In *Behavioral Simulation Methods in Tax Policy Analysis*, edited by Martin Feldstein. Chicago: Univ. Chicago Press (for N.B.E.R.), 1983.

Brittain, John A. *Inheritance and the Inequality of Material Wealth*. Washington Brookings Inst., 1978.

Kotlikoff, Laurence J., and Summers, Lawrence H. "The Role of Intergenerational Transfers in Aggregate Capital Accumulation." *J.P.E.* 89 (August 1981): 706–32

Mirer, Thad W. "The Wealth-Age Relation among the Aged." *A E R* 69 (June 1979): 435–43

Seidman, Laurence S. "The Personal Consumption Tax and Social Welfare" *Challenge* 23 (September 1980): 10–16.

———. "Conversion to a Consumption Tax: The Transition in a Life Cycle Growth Model." Working Paper. Newark: Univ. Delaware, 1983. (*a*)

———. "Taxes in a Life Cycle Growth Model with Bequests and Inheritances." *A.E.R.* 73 (June 1983): 437–41. (*b*)

Seidman, Laurence S., and Maurer, Stephen B. "Taxes and Capital Intensity in a Two-Class Disposable Income Growth Model." *J. Public Econ* 19 (November 1982): 243–59.

Strotz, Robert H. "Myopia and Inconsistency in Dynamic Utility Maximization." *Rev. Econ. Studies* 23, no. 3 (1956): 165–80.

Summers, Lawrence H. "Tax Policy in a Life Cycle Model." Working Paper no. 302. Cambridge, Mass.: N.B.E.R., 1979. Shortened version ("Capital Taxation and Accumulation in a Life Cycle Growth Model") in *A E R* 71 (September 1981): 533–44.

White, Betsy B. "Empirical Tests of the Life Cycle Hypothesis." *A.E.R.* 68 (September 1978): 547–60.

Aggregate Information and the Role of Monetary Policy in an Open Economy

Kent P. Kimbrough

Duke University

A model of a small open economy in which agents trade in local goods markets and an economy-wide asset market is developed. Purchasing-power parity is assumed to hold at the aggregate level. However, because of local deviations from purchasing-power parity, agents possess differential information. Using this framework, it is shown that when the exchange rate is flexible monetary policy can influence the distribution of real output by altering the information content of the exchange rate. However, when monetary policy is committed to fixing the exchange rate (by a feedback rule) the distribution of real output is independent of the particular exchange rate rule chosen. The stability of real output under the two regimes is compared, and it is demonstrated that regardless of the stability of domestic monetary policy a flexible exchange rate regime is superior in this respect. Possible qualifications and extensions of these results are also discussed.

The policy implications of the new classical macroeconomics have attracted considerable attention. The most controversial result has been the policy-ineffectiveness proposition advanced by Sargent and Wallace (1975). Recently papers by Weiss (1980, 1982) and King (1982) have demonstrated that this policy-ineffectiveness proposition breaks down if agents with differential information transact in a common market and thus have access to a common price. Their

---

I would like to thank Grant Gardner, George Tauchen, and two anonymous referees for their helpful comments. See King (1983) for a closed-economy treatment of closely related issues. King's paper came to my attention at too late a stage to be incorporated into the body of the paper.

[*Journal of Political Economy*, 1984, vol. 92, no. 2]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0005$01.50

arguments hinge on the ability of monetary policy to influence the information content of prices that serve to aggregate the diverse information of market participants. The aim of this paper is to extend to the open economy some of these ideas and policy implications that have recently emerged from the literature on the new classical macroeconomics.

The model employed in the paper, and developed in Section I, is an open-economy version of the "island" paradigm popularized by Lucas (1972, 1973), which is intended to capture the essential elements of the incomplete information view of the business cycle advanced by Friedman (1968) and Phelps (1970). Agents trade in spatially separated goods markets, with prices differing across markets because agents in each market choose different foreign trading partners at the start of each period. Purchasing-power parity holds at the aggregate level, but there are local (market-specific) deviations. In addition, agents also trade in integrated asset markets and thus have access to common price information, in particular the exchange rate and foreign interest rate. The incorporation of an economy-wide asset market into the model makes it similar in spirit to Barro (1980).[1] Since agents with diverse information are trading in a common market, the framework is well suited for discussing the policy-ineffectiveness proposition and when it may break down. This discussion is the subject of Sections II–IV. Sections II and III below examine the determination of equilibrium and the role of monetary policy under a flexible exchange rate regime. Section IV examines the role of exchange rate policy under a fixed exchange rate regime. Finally, Section V contrasts the role of macroeconomic policy under fixed and flexible exchange rate regimes and compares the amplitude of the business cycle under the two regimes. Possible qualifications and extensions are discussed.

Before proceeding, it will be useful to summarize briefly the conclusions of the paper. First, it is demonstrated that under flexible exchange rates systematic monetary policy can influence the distribution of real output by altering the information content of the exchange rate. This is an open-economy extension of the results derived by Weiss (1980, 1982) and King (1982). Second, under fixed exchange rates the monetary authorities cannot use exchange rate policy to influence the distribution of real output, because exchange rates fail to convey useful information. Third, it is shown that for a small open economy flexible exchange rates are superior to fixed exchange rates, in the sense that they result in a smaller variance of real output about its full current information value, regardless of the wisdom or discipline with which monetary policy is conducted.

---

[1] Harris and Purvis (1981) use a model similar to the one developed here to examine exchange rate behavior and market efficiency

I. The Model

The country under consideration is assumed to be a small open economy. One good is produced and consumed by domestic agents who are located in spatially separated markets indexed by the letter z; there is no trade in goods across markets at a point in time. The distinguishing characteristic of each market is its choice of a foreign trading partner at the start of each period. Because different domestic markets have different foreign trading partners, commodity prices differ from market to market as specified below. However, at the aggregate level purchasing-power parity holds—deviations from purchasing-power parity are strictly localized and uncorrelated over time. The only assets domestic agents are assumed to hold are domestic money and foreign bonds. Asset markets are integrated so that agents in all markets observe the nominal interest rate on foreign bonds, i^*, and the (natural logarithm of the) exchange rate, S_t, defined as the price of foreign exchange in terms of domestic currency. Agents in market z also know (the natural logarithm of) their local commodity prices measured in home and foreign currencies, $P_t(z)$ and $P_t^*(z)$.

The formal structure of the model is summarized by the following (all variables are in natural logarithms except for i^* and i_t, which are in levels):

$$P_t(z) = S_t + P_t^*(z), \tag{1}$$

$$P_t^*(z) = P_t^* + z_t, \tag{2}$$

$$P_t^* = P_{t-1}^* + v_t, \tag{3}$$

$$y_t(z) = \alpha r_t(z) + u_t; \; r_t(z) = P_t^*(z) - E_z P_{t+1}^* + i^*, \tag{4}$$

$$M_t^d = P_t + \phi y_t - \lambda i_t + w_t. \tag{5}$$

Expression (1) is a goods-market arbitrage condition for market z. As shown by (2), goods prices in the local community's foreign trading partner, $P_t^*(z)$, deviate from the economy-wide average foreign price, P_t^*, by an amount z_t, where $z_t \sim N(0, \sigma_z^2)$. Foreign prices differ across markets because the foreign country is also assumed to consist of a set of local markets. Since the home country is small, domestic agents take foreign prices and their distribution as given. If (1) and (2) are combined, it can be seen that $P_t(z) = S_t + P_t^* + z_t$, which implies that the economy-wide average of local commodity prices (the price level) is $P_t = S_t + P_t^*$. Therefore, purchasing-power parity holds at the aggregate level, but there are local deviations reflected by z_t. As shown by (3), it is assumed for simplicity that the stochastic structure of the foreign economy is such that the economy-wide average foreign price follows a random walk with $v_t \sim N(0, \sigma_v^2)$.

The local commodity supply function (4) captures the intertemporal aspects of the model. An increase in the real rate of return, or relative price of current goods in terms of future goods, perceived by agents in market z, $r_t(z)$, causes them to substitute future for current leisure and thus results in an increase in current real output, $y_t(z)$. The relative price term can be explained as follows: a sale this period allows agents to accumulate $P_t^*(z)$ plus the nominal rate of return on foreign bonds by period $t + 1$, while a sale next period is expected to yield $E_z P_{t+1}^*$, which is the expected future foreign price conditioned on the information available to agents in market z. Comparison of these yields results in the relative price term $P_t^*(z) - E_z P_{t+1}^* + i^*$. Real output is also influenced by an aggregate supply shock, u_t, which is normally and independently distributed with mean zero and variance σ_u^2.

The aggregate demand for money, M_t^d, is given by (5) and depends on the economy-wide average levels of local prices, real output, and the perceived opportunity cost of holding money, which are given by P_t, y_t, and i_t. In addition, there is also an aggregate money demand shock, $w_t \sim N(0, \sigma_w^2)$. This specification follows from the fact that asset markets are integrated and can be viewed as having been obtained b averaging local money demand functions of the same fo markets. The opportunity cost of holding money as perc agents in market z is the forgone interest earnings on th bond plus the expected rate of depreciation of the domestic $E_z S_{t+1} - S_t$. Averaging this opportunity cost across market that

$$i_t = i^* + (\overline{E_z S_{t+1}} - S_t), \tag{6}$$

where $\overline{E_z S_{t+1}}$ is the economy-wide average of local expectations concerning the future exchange rate, $E_z S_{t+1}$.

II. Flexible Exchange Rates and Full Current Information

Under a flexible exchange rate regime the domestic money supply is a policy variable that is determined by the monetary authorities, and the exchange rate adjusts to maintain asset market equilibrium. In this and the following section the money supply, M_t, is assumed to be governed by the rule

$$M_t = M_{t-1} + m_t + \psi v_{t-1}, \tag{7}$$

where ψ is a policy parameter chosen by the monetary authorities, and the nonsystematic element of monetary policy, m_t, is assumed normally and independently distributed with mean zero and variance σ_m^2. The policy rule (7) can easily be altered to allow for a response to other lagged shocks without affecting any of the paper's qualitative conclusions.

A. *Solution Technique*

The most straightforward way to solve for the equilibrium under flexible exchange rates is to aggregate and then use the method of undetermined coefficients to solve for the equilibrium exchange rate.

Substitution of (1)–(3) into the relative price term, $r_t(z)$, yields

$$r_t(z) = i^* + z_t + (v_t - E_z v_t). \tag{8}$$

Using (8) in connection with the local supply function (4) when averaging across markets gives an expression for aggregate real output, y_t. When the resulting expression for aggregate real output is combined with (5)–(7) and the expression for the price level, P_t, that emerges from (1)–(3), it can be shown that asset market equilibrium requires

$$\begin{aligned}(1 + \lambda)S_t = {} & M_{t-1} + m_t + \psi v_{t-1} - P^*_{t-1} - (1 + \phi\alpha)v_t - \phi u_t - w_t \\ & + (\lambda - \phi\alpha)i^* + \phi\alpha\overline{E_z v_t} + \lambda\overline{E_z S_{t+1}},\end{aligned} \tag{9}$$

where $\overline{E_z v_t}$ is the economy-wide average of $E_z v_t$.

Applying the method of undetermined coefficients amounts to "guessing" a solution for the equilibrium exchange rate of the form[2]

$$\begin{aligned}S_t = {} & \pi_0 + \pi_1 M_{t-1} + \pi_2 P^*_{t-1} + \pi_3 v_{t-1} \\ & + \pi_4 m_t + \pi_5 v_t + \pi_6 u_t + \pi_7 w_t.\end{aligned} \tag{10}$$

The assumption of rational expectations requires that the expected future exchange rate be found by updating (10) one period and taking expectations conditional on all available information (which depends on the market in which an agent is currently located). With (7) and (9) this procedure yields

$$\begin{aligned}E_z S_{t+1} = {} & \pi_0 + \pi_1(M_{t-1} + E_z m_t + \psi v_{t-1}) \\ & + \pi_2(P^*_{t-1} + E_z v_t) + \pi_3 E_z v_t.\end{aligned} \tag{11}$$

---

[2] Note that local deviations from purchasing-power parity do not influence the exchange rate, since it is determined in an economy-wide asset market.

Finally, the π coefficients can be found from (9), (10), and the economy-wide average of (11).

B. Equilibrium with Full Current Information

When agents possess full current information, $E_z m_t = m_t$ and $E_z v_t = v_t$. From (11) this can be seen to imply that $E_z S_{t+1} = \overline{E_z S_{t+1}}$. With these facts, solutions for the full current information π's (denoted by a tilde) can be obtained. These solutions are $\tilde{\pi}_0 = (\lambda - \phi\alpha)i^*$, $\tilde{\pi}_1 = \tilde{\pi}_1 = 1$, $\tilde{\pi}_2 = -1$, $\tilde{\pi}_3 = \psi$, $\tilde{\pi}_5 = -1 + [\lambda\psi/(1 + \lambda)]$, $\tilde{\pi}_6 = -\phi/(1 + \lambda)$, and $\tilde{\pi}_7 = -1/(1 + \lambda)$.

Regarding the real equilibrium of the economy, it can be seen from (8) that under full current information the real rate of return perceived by agents in market z is

$$\tilde{r}_t(z) = \imath^* + z_t. \tag{12}$$

With this in (4), it follows that the full current information level of real output in market z is

$$\tilde{y}_t(z) = \alpha(\imath^* + z_t) + u_t. \tag{13}$$

As can be seen from this expression, a positive value of z_t causes agents in market z to increase their real output relative to the aggregate level, $\tilde{y}_t = \alpha\imath^* + u_t$. The reason for this is that agents in market z are aware that the real rate of return in their market is unusually high, and in order to take advantage of this opportunity they substitute future for current leisure. This expansion in current labor supply results in an increase in real output in market z.

III. Flexible Exchange Rates and Incomplete Information

In the previous section the benchmark case of full current information was considered. In this section it is assumed that agents must use the market prices they observe to make inferences about the state of the economy. These prices will not, in general, perfectly reveal the state of the economy, and agents will thus have to make their supply and demand decisions on the basis of incomplete information. Equations (8)–(11) of Section II*A* continue to hold in this instance. The main difference is that in solving the model $E_z m_t$ and $E_z v_t$ cannot simply be replaced by their actual values as under full current information.

A. *Extracting Information from Prices*

In the model employed here agents observe four market prices—the nominal interest rate on foreign bonds and the exchange rate, both economy-wide prices, and their own local commodity prices, $P_t(z)$ and $P_t^*(z)$. Only two of these prices, the exchange rate and the local foreign currency commodity price, are useful to agents in their attempt to assess current economic conditions. This follows from the fact that, in the interest of parsimony, the nominal interest rate on foreign bonds is assumed to be uncorrelated with any contemporaneous shocks and that $P_t(z)$ is redundant given the arbitrage condition (1).[3]

The problem confronting agents is to use the information at their disposal to form conditional expectations of the current shocks. Equation (10) indicates that the information content of the exchange rate is $\pi_4 m_t + \pi_5 v_t + \pi_6 u_t + \pi_7 w_t$, and from (2) and (3) it follows that the information content of $P_t^*(z)$ is $v_t + z_t$. From (11) it can be seen that a solution to the model requires explicit calculation of $E_z m_t$ and $E_z v_t$ only. These expectations are given by

$$E_z m_t = \theta_1(\pi_4 m_t + \pi_5 v_t + \pi_6 u_t + \pi_7 w_t) + \theta_2(v_t + z_t), \quad (14)$$

$$E_z v_t = \beta_1(\pi_4 m_t + \pi_5 v_t + \pi_6 u_t + \pi_7 w_t) + \beta_2(v_t + z_t), \quad (15)$$

where

$$\theta_1 = \frac{\pi_4 \sigma_m^2(\sigma_v^2 + \sigma_z^2)}{\Delta}, \quad \theta_2 = \frac{-\pi_4 \pi_5 \sigma_m^2 \sigma_v^2}{\Delta},$$

$$\beta_1 = \frac{\pi_5 \sigma_v^2 \sigma_z^2}{\Delta}, \quad \beta_2 = \frac{\sigma_v^2(\pi_4^2 \sigma_m^2 + \pi_6^2 \sigma_u^2 + \pi_7^2 \sigma_w^2)}{\Delta},$$

and

$$\Delta = (\sigma_v^2 + \sigma_z^2)(\pi_4^2 \sigma_m^2 + \pi_6^2 \sigma_u^2 + \pi_7^2 \sigma_w^2) + \pi_5^2 \sigma_v^2 \sigma_z^2 > 0.$$

B. *Equilibrium with Incomplete Information*

The first step in providing a complete solution to the model would be to use the method of undetermined coefficients to solve for the π's, θ's, and β's. However, this paper is concerned primarily with the behavior of real variables, and it turns out that quite a few conclusions can be drawn without solving explicitly for these coefficients.[4] For the discussion that follows it is sufficient to note that $\pi_i = \bar{\pi}_i$ for $i = 0, \ldots, 3$, since agents continue to know the economy's past history.

---

[3] It would be possible, without altering any conclusions, to extend the model to allow for the foreign nominal interest rate to be correlated with current shocks, and in particular those influencing economy-wide foreign prices. This could be done along the lines of Flood and Marion (1982).

[4] About the best that can be done in this respect is to substitute out for the θ's and β's

In order to examine the real equilibrium, it is convenient to solve the model in terms of the deviations of the incomplete information values of local relative prices and real outputs from their full current information values. From (8) and (12) it can be seen that

$$r_t(z) - \bar{r}_t(z) = v_t - E_z v_t, \tag{16}$$

and from (4) it follows that

$$y_t(z) - \bar{y}_t(z) = \alpha(v_t - E_z v_t). \tag{17}$$

Therefore, deviations of real variables from their full current information values arise solely from the errors agents make in predicting the economy-wide element of the contemporaneous foreign price shock, v_t. Since agents observe $P_t^*(z) = v_t + z_t$, expectational consistency requires that $E_z v_t + E_z z_t = v_t + z_t$, which implies that $v_t - E_z v_t = E_z z_t - z_t$. Hence if agents in market z underestimate the economy-wide component of the foreign price shock, they will overestimate the local component by the same amount and mistakenly perceive the local real rate of return to be unusually high, as indicated by (16). This will cause agents to increase their current labor supply and cause real output to exceed its full current information value, as shown by (17). A similar story can be told if agents overestimate the economy-wide component of the foreign price shock.

In order to determine how agents respond to various current shocks, note from (15) that

$$v_t - E_z v_t = (1 - \beta_1 \pi_5 - \beta_2) v_t - \beta_1 (\pi_1 m_t + \pi_6 u_t + \pi_7 w_t)$$

It can be shown from the expressions for β_1 and β_2 presented
and from the fact that $\pi_5 = -[1 + \phi\alpha - \lambda\theta_2 + (\lambda - \phi\alpha - \lambda$
$+ \lambda(1 - \theta_1) + (\lambda - \phi\alpha - \lambda\psi)\beta_1]$ that $1 - \beta_1 \pi_5 - \beta_2 = \sigma$
$\pi_6^2 \sigma_u^2 + \pi_7^2 \sigma_w^2)/\Delta \geqslant 0$ and that $\beta_2 \geqslant 0$. Therefore, as can be se
(17) and (18), increases in the economy-wide component o
foreign price shock, v_t, raise real output relative to its full current information value, while increases in the local component, z_t, reduce real output relative to its full current information value. That is, real output responds "too much" to aggregate shocks and "not enough" to local (or relative) shocks. The reason for this is that whenever a posi-

---

in the expressions for π_1, π_5, π_6, and π_7 The solutions for these coefficients will then be implicit in the resulting polynomial equations. However, few insights can be gained from this tedious exercise, the results of which are available from the author on request This difficulty is a general feature of macroeconomic models with differential information such as the one presented here (see, e g , Barro 1980; King 1982) tive aggregate foreign price shock occurs, agents mistakenly perceive part of the change to be the result of a positive local shock, in which case they would like to expand current real output to take advantage of the high local real rate of return. A similar story explains the effects of local shocks.[5]

C. *Effectiveness of Monetary Policy*

A key feature of the model presented here is that the policy rule chosen by the monetary authorities can influence the distribution of real output (and the perceived real rate of return). It can do so because agents have differential information but transact in a common market, and hence there is room for monetary policy to be effective by altering the information content of commonly observed market prices. In this case it is the information content of the exchange rate that can be altered by monetary policy. Insight into this result can be obtained by noting from (11) and the solutions for π_0, π_1, π_2, and π_3 that

$$\overline{E_z S_{t+1}} = S_t^0 + \overline{E_z m_t} + (\psi - 1)\overline{E_z v_t},$$

where $S_t^0 = \pi_0 + \pi_1(M_{t-1} + \psi v_{t-1}) + \pi P_{t-1}^*$, is that part of the exchange rate that can be predicted on the basis of past information. Substituting this into (9) and rearranging yields

$$\begin{aligned}(1 + \lambda)(S_t - S_t^0) = {} & [m_t - (1 + \phi\alpha)v_t - \phi u_t - w_t] \\ & + \{\lambda\overline{E_z m_t} + [\phi\alpha + \lambda(\psi - 1)]\overline{E_z v_t}\}.\end{aligned} \tag{19}$$

Since S_t^0 is known to all agents from their knowledge of the economy's past history, the right-hand side of (19) describes the new information conveyed to agents by the current exchange rate. This new information has two distinct components: the first concerns the combination of aggregate shocks $m_t - (1 + \phi\alpha)v_t - \phi u_t - w_t$, and the second market expectations as reflected by the term $\lambda\overline{E_z m_t} + [\phi\alpha + \lambda(\psi - 1)]\overline{E_z v_t}$. If agents in different markets are differentially informed they will not know the market expectations component of (19), and because of this noise they will form different expectations of the aggregate shocks component, $m_t - (1 + \phi\alpha)v_t - \phi u_t - w_t$. Monetary policy can alter the distribution of market expectations through its impact on the information content of the exchange rate, which arises because the policy parameter ψ enters the second term of (19). The policy parameter ψ enters (19) and thus influences the information content of the exchange rate, both directly and indirectly, through its effect on the endogenous expectations $\overline{E_z m_t}$ and $\overline{E_z v_t}$. Because monetary policy can alter the information content of the exchange rate, it can influence the distribution of real output about its full current information value since this distribution depends on agents' forecast errors.[6]

---

[5] Aggregate demand for this economy will have the same form as the aggregate supply function obtained by averaging (4) across all markets, except that it will be inversely related to the perceived real rate of return. Therefore, under flexible exchange rates the trade account will respond "too much" to aggregate shocks—the capital account will behave similarly (since the two must add to zero). Furthermore, the model predicts that the trade-account surplus will be positively correlated with real output over the business cycle. Under fixed exchange rates similar results hold for the trade balance, but its movements need not be mirrored by the capital account because of fluctuations in the balance of payments. These remarks assume that wealth effects are small. See Stockman and Koh (1982) for a discussion that includes wealth effects.

In the case where there are no local deviations from purchasing-power parity, agents are not differentially informed and local and market expectations are identical (i.e., $E_z m_t = \overline{E_z m_t}$, etc.). Therefore, monetary policy cannot alter the distribution of real output since all agents know market expectations and can thus reduce the information content of the exchange rate to its aggregate shocks component, $m_t - (1 + \phi\alpha)v_t - \phi u_t - w_t$, which is independent of the policy parameter ψ.[7] However, agents still have incomplete information since they cannot accurately predict the four remaining shocks (m_t, u_t, v_t, and w_t) from the knowledge of the exchange rate and local foreign currency prices, $P_t^*(z)$. This means that monetary policy does not attain potency simply because agents lack full current information. On the other hand, when local deviations from purchasing-power parity are present, monetary policy is able to influence the distribution of real output. But in this case, agents not only have incomplete information—they also have differential information [illegible] shown by (14) and (15), the market in which they are curr[illegible] colors their view of the state of the economy. The ups[illegible] monetary policy is able to influence the distribution of re[illegible] only to the extent that agents transacting in a common m[illegible] incomplete and differential information.

Finally, to see that monetary policy gains its leverage because [illegible] ability to alter the information content of the exchange rate, not of local prices, note that the recursive projection formula implies that

$$E_z v_t = E_z[v_t|P_t^*(z)] + E_z\{v_t - E_z[v_t|P_t^*(z)]|S_t - S_t^0 - E_z[S_t - S_t^0|P_t^*(z)]\}. \tag{20}$$

---

[6] If the policy rule (7) were extended to include other lagged shocks, the policy parameters attached to them would also influence the distribution of real output. Flood and Hodrick (1982) have shown that feedback on the current exchange rate is irrelevant for the distribution of real output, since it is observed by private agents.

[7] When there are no local deviations from purchasing-power parity $s_t(z) = \hat{s}_t(z)$ in the setup adopted here since knowledge of $P_t^*(z)$ fully reveals v_t. If foreign prices were subject to both permanent and transitory shocks this extreme result would not hold.

which follows from the fact that agents observe $P_t^*(z)$ and $S_t - S_t^0$. It can easily be demonstrated that

$$E_z[v_t|P_t^*(z)] = \gamma(v_t + z_t); \quad \gamma = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_z^2}. \tag{21}$$

Since the first term of (20) is, therefore, independent of the money supply rule, it follows that monetary policy influences agents' prediction errors, and hence the distribution of real output, because of its ability to influence the information content of the exchange rate as reflected by the second term of (20). This has important implications for the comparison of the variance of real output under fixed and flexible exchange rates that is taken up in Section V. Before turning to this issue, however, the next section will consider the case of a fixed exchange rate regime.

IV. Fixed Exchange Rates

Under a fixed exchange rate regime, which here means any system in which the exchange rate is set by a feedback rule, the monetary authorities sacrifice control of the money supply in order to manipulate the exchange rate. For simplicity, it is assumed that the monetary authorities set the exchange rate according to the rule

$$S_t = S_{t-1} + \rho v_{t-1}, \tag{22}$$

where ρ is a policy parameter reflecting the extent to which exchange rate changes are used to offset ($\rho < 0$) or reinforce the price-level effects of past foreign price shocks. This particular rule has been chosen to parallel the money supply rule (7) that was used in the discussion of flexible exchange rates.[8] The rule can easily be generalized to allow for a policy response to other lagged shocks without altering any of the conclusions of this section.

Because the monetary authorities are committed to buy and sell foreign exchange in order to support the exchange rate rule (22), movements in the money supply replace the exchange rate as the mechanism that equilibrates the asset market. From (22) it can be seen that $\overline{E_z S_{t+1}} = S_t + \rho \overline{E_z v_t}$, and with this and (22) in the asset market equilibrium condition (9), it can be shown that the money supply must satisfy the condition

$$\begin{aligned} M_t = {} & S_{t-1} + \rho v_{t-1} + P_{t-1}^* + (1 + \phi\alpha)v_t + \phi u_t + w_t \\ & - (\lambda - \phi\alpha)i^* - (\phi\alpha + \lambda\rho)\overline{E_z v_t}. \end{aligned} \tag{23}$$

---

[8] No random element comparable to m_t is included in (22) since it would be revealed to agents via their observation of the current exchange rate.

Equation (23) can be used to solve for the equilibrium money supply under full current information and incomplete information. However, as before, this is not done explicitly at this point, since it is the behavior of real rather than nominal variables that is of primary concern.

As in the flexible exchange rate case, it is still true that deviations of real variables from their full current information values arise only to the extent that agents make errors in predicting economy-wide foreign price shocks. Therefore, in the case of incomplete information, (16) and (17) continue to hold under fixed exchange rates. To solve for the real equilibrium of the model under incomplete information and a fixed exchange rate, an expression for $E_z v_t$ is needed. As before, the recursive projection formula implies that this expectation is given by (20). However, since the exchange rate is determined by the feedback rule (22), it follows that $S_t \equiv S_t^0$ and that the second term in (20) is identically zero. Therefore, under a fixed exchange rate regime the conditional expectation of the economy-wide foreign price shock held by agents in market z is given by (21). Substitution of (21) into (16) and (17) gives $r_t(z) - \bar{r}_t(z) = (1 - \gamma)v_t - \gamma z_t$ and

$$y_t(z) - \bar{y}_t(z) = \alpha(1 - \gamma)v_t - \alpha\gamma z_t. \qquad (24)$$

As in the flexible exchange rate case, real output responds "too much" to aggregate foreign price shocks and "not enough" to local shocks.

The major implication of (24) is that the exchange rate rule does not affect the distribution of real output even though agents h
incomplete and differential information.[9] The reason for t
when the exchange rate is a policy variable determined by
rule, the monetary authorities cannot influence its inform
tent. Only to the extent that the exchange rate responds
conditions, as it does under a flexible exchange rate regim
icy rules possibly influence the distribution of real output
exchange rate by a feedback rule robs the exchange rate of its
mation content and, for better or worse, deprives the monetary authorities of any power they might otherwise have had to affect the distribution of real output.

V. A Comparison of Alternative Exchange Rate Regimes

In a comparison of fixed and flexible exchange rate regimes two major results emerge. These results are presented and then possible qualifications and extensions are discussed.

---

[9] Note that the policy parameter ρ does not appear in (24)

A. *The Role of Policy and the Superiority of a Flexible Rate*

The first major result concerns the role of monetary policy under fixed and flexible exchange rates. When the exchange rate is flexible the monetary authorities can alter the distribution of real output by the money supply rule that they choose as shown in Section IIIC. The effectiveness of monetary policy in this case arises from its ability to alter the information content of the exchange rate when agents possess incomplete and differential information.[10] In contrast, under a fixed exchange rate regime the monetary authorities are powerless. Exchange rate policy cannot influence the distribution of real output since setting the exchange rate by a feedback rule entirely destroys its information content, thus shutting off the channel through which monetary policy gains effectiveness. Therefore, when a small open economy moves away from a flexible exchange rate toward a fixed exchange rate, the monetary authorities lose any ability to influence the distribution of real output: it is not simply the case that monetary policy works through different channels; rather it works in one case and not in the other. This result can be viewed as a rational expectations counterpart to Mundell's (1963) results concerning the effectiveness of monetary policy under fixed and flexible exchange rates.

Second, it can be shown that a flexible exchange rate regime unambiguously results in a smaller variance of real output about its full current information value than does a fixed exchange rate regime.[11] That is, a small open economy with a fixed exchange rate can reduce the amplitude of its business cycle by adopting a flexible exchange rate. To see this, note from (17) that the variance of real output about its full current information value is directly related to the variance of agents' forecast errors in predicting the economy-wide component of the foreign price shock, v_t. When the recursive projection formula is applied, this forecast is given by (20). Under both regimes the first term of (20) is given by (21). However, under a fixed exchange rate the second term in (20) is zero, while under flexible exchange rates it is not. This reflects the fact that when the exchange rate is flexible it conveys useful information, but when it is fixed by a feedback rule it does not. Therefore, because a flexible exchange rate provides agents with more information than does a fixed exchange rate, and since more information always improves (or at least does not worsen) fore-

---

[10] The ability of monetary policy to influence the distribution of real output in the model presented here is in contrast to results derived by Kimbrough (1983*a*). The difference arises because it is assumed there that all agents possess the same information and hence no differential information exists for policymakers to exploit.

[11] Flood and Hodrick (1982) reach the same conclusion for similar reasons using a different model.

casting accuracy, it follows that adoption of a flexible exchange rate will result in a smaller variance of real output about its full current information value. In the model presented here, this is true except in the limiting case where the variance of the new information contained in the exchange rate conditional on the information conveyed by the local foreign price, var $[S_t - S_t^0 | P_t^*(z)]$, approaches infinity, since observing a totally noisy signal provides no information at all. This could be the case if, for instance, the monetary authorities refuse to follow the systematic part of their policy rule very closely and $\sigma_m^2 \rightarrow \infty$. In this case, the variance of real output is the same under the two regimes. Hence one cannot argue that a fixed exchange rate is to be preferred on the grounds that it "disciplines" monetary policy by committing the monetary authorities to peg the exchange rate. Regardless of the wisdom (in the sense of making judicious choices of policy parameters such as ψ) or discipline with which monetary policy is conducted, for a small open economy the amplitude of the business cycle should be less with a flexible exchange rate than with a fixed exchange rate.[12] This "superiority" of flexible exchange rates arises solely because of the power of market-determined prices to convey useful information to rational agents—a power that is forfeited when policymakers attempt to fix the exchange rate by a feedback rule. It is important to note that the superiority of flexible exchange rates holds even though money demand shocks influence domestic prices under floating rates (via the exchange rate), while domestic prices are insulated from money demand shocks (of domestic origin) under fixed exchange rates. The reason that this is irrelevant for the relative stability of real
under the two regimes is that a small country takes the
return, $r_t(z)$, which matters for output decisions as giv
while the two regimes differ in terms of the impact of dom
demand shocks on nominal prices, intertemporal relativ
small country are insulated from domestic money dem
under both regimes.

B. Qualifications and Extensions

There appear to be five important qualifications and possible extensions of the results presented in this paper. These qualifications and extensions take three basic forms—altering the information set.

---

[12] Of course, wisdom and discipline in the conduct of monetary policy are desirable. the argument here is not to the contrary The claim is that a small open economy can reduce the amplitude of its business cycle by adopting a flexible exchange rate. Then, having adopted a flexible exchange rate, policymakers can concern themselves with finding the optimal money supply rule. In the context of the model presented here this involves choosing the values for ψ and σ_m^2 that minimize the variance of real output about its full current information value.

changing the specification of the local commodity supply function, and respecifying or extending the markets in which agents are assumed to transact.

The first qualification concerns the information set agents possess. The main extension in this regard is that agents could be assumed to observe monetary aggregates contemporaneously, rather than with a lag as in the present setup. This information could be either exact or noisy, as in King (1981). In either case, the results would be modified in two ways. First, exchange rate policy would be able to affect the distribution of real output by altering the information content of the observed monetary aggregate. This issue is taken up by Kimbrough (1983*b*). Second, the "superiority" of a flexible exchange rate regime would be called into question because, although it is still true that the information content of the exchange rate is nil under a fixed exchange rate regime, the information conveyed by observing monetary aggregates might be enough to offset this disadvantage.[13] One difficulty with this extension, as noted by King (1981), is that if agents observe monetary aggregates contemporaneously, then an econometrician with data on these aggregates should find them to be uncorrelated with real output. However, empirical evidence clearly supports the existence of a positive correlation between money and real output over the course of the business cycle. This extension must, therefore, be reconciled with the empirical evidence before it can be accepted as valid. Another extension in the same spirit would be to assume that agents know their own money holdings, $M_t(z)$, rather than the aggregate money supply, M_t. As shown in an appendix that is available on request, this can be viewed as a special case of noisy information about the aggregate money supply. However, it can also be shown that expanding the information set in this manner does not alter the paper's qualitative conclusions concerning the role of monetary policy under fixed and flexible exchange rates and the "superiority" of flexible exchange rates. In fact, the availability of information about *local* money holdings actually strengthens the case for flexible exchange rates.[14]

---

[13] To compare fixed and flexible exchange rate regimes in this case simulation methods would be needed to (i) find optimal money supply and exchange rate rules and (ii) compare the stability of real output under the two regimes when the optimal rule is in force. The need for simulation stems from the fact that a closed-form solution to the model is not readily obtained analytically.

[14] Under both fixed and flexible exchange rates, agents observe local foreign currency prices and economy-wide interest rates. Since local money holdings depend on these variables and since foreign prices and interest rates are not influenced by domestic shocks in the small country case, it follows that agents can reduce the information content of local money holdings to a combination of domestic shocks that is uncorrelated with foreign prices and interest rates. Therefore, under fixed exchange rates

A second extension concerns adding a wealth term to the local commodity supply function along the lines of Barro (1980). In this case, monetary discipline (as measured by σ_m^2) is an important factor in comparing exchange rate regimes (provided that wealth effects are not very weak). It can be shown, however, that as long as monetary policy under flexible exchange rates is not too undisciplined (i.e., as long as σ_m^2 is not too large), flexible exchange rates will be superior to fixed exchange rates in stabilizing real output. The results presented earlier can be viewed as being based on the assumption that wealth effects are weak, as reported by Lucas and Rapping (1969).

Yet a third qualification concerns extensions of the model designed to make the country large in some markets. This could be done by making the country large in world markets (say small in its import market and large in its export market) or by introducing into the model nontraded goods or assets that are traded economy wide. With these modifications, exchange rate policy could influence the distribution of real output by altering the information content of an endogenous price other than the exchange rate, so long as that price is observed by all agents. In addition, no unambiguous conclusion regarding the amplitude of the business cycle under fixed and flexible exchange rates would emerge in this case. This result is due to differences in the stochastic structure of the economy under the two regimes, as shown by Kimbrough (1983a). These qualifications serve to highlight the crucial role played by parity conditions in deriving the results presented earlier.

A fourth qualification concerns private indexing based on cu local commodity prices or the exchange rate. As shown (1982), such indexing schemes are comparable to moneta that they can affect the distribution of real output thro fluence on the information content of market prices. I such an indexing scheme into the present setup would i implications along the lines suggested by King (1982). Ho superiority of a flexible exchange rate in terms of sta output would remain intact, since it would still be true th change rate conveys useful information when it is flexible bu when it is fixed.

Finally, it should be noted that the model presented here does not

---

the information provided by local money holdings will not help agents to forecast v. However, under flexible exchange rates agents have two bits of information, local foreign prices and the exchange rate, that are useful in forming forecasts of v_n, and the exchange rate is contaminated by various domestic shocks. Hence the availability of local money supply information will improve agents' forecasts of v_n because they can use the knowledge about domestic shocks it provides to "sharpen" the information about the aggregate foreign price shock, v_n, conveyed to them by the exchange rate.

include a forward market in foreign exchange. In Kimbrough (1982) the model is extended in this direction. There, it is shown that the properties of the forward rate as a predictor of the future spot rate are a crucial determinant of the effectiveness or ineffectiveness of monetary and exchange rate policies and of the relative stability of real output under fixed and flexible exchange rates. The omission of a forward market in the present model can be viewed in one of three ways. First, it can be viewed as an assumption that for the currency of the country in question no such market exists, either because of legal or other institutional constraints or simply because the market is "too thin." Second, empirical evidence indicates that movements of spot and forward exchange rates are highly correlated (see Frenkel 1981). This suggests that spot and forward rates convey pretty much the same information. Therefore, as a practical matter, the omission of a forward market may not be too serious. Third, the model presented here can be viewed as making the extreme, but simplifying, assumption that the forward rate is a totally noisy signal and thus conveys no useful information (although empirical evidence would certainly contradict this assumption).

Now that earlier results have been qualified in several respects, it is important to reassess them in the light of these qualifications. Regarding the role of monetary policy under flexible exchange rates, none of the qualifications just noted alters the conclusion that it can influence the distribution of real output so long as agents with incomplete and differential information trade in a common market. However, to the extent that other endogenous variables besides the exchange rate, be they prices or monetary aggregates, are observed by all agents, the conclusion that exchange rate policy is ineffective may no longer hold. Finally, the superiority of a flexible exchange rate in stabilizing real output is subject to some qualifications. However, for a country that is small in world goods markets and whose nontraded goods and assets are close substitutes for traded goods and assets, so that their prices are closely aligned with world prices, adopting a flexible exchange rate is likely to reduce the amplitude of the business cycle (if wealth effects are weak and/or monetary policy is disciplined). This result highlights the role of market-determined prices in facilitating the efficient allocation of scarce resources by conveying useful information to rational agents.

References

Barro, Robert J. "A Capital Market in an Equilibrium Business Cycle Model." *Econometrica* 48 (September 1980): 1393–1417.

Flood, Robert P., and Hodrick, Robert J. "Optimal Price and Inventory Adjustment in an Open-Economy Model of the Business Cycle." Unpublished manuscript. Washington: Bd. Governors Fed. Reserve System; and Pittsburgh: Carnegie-Mellon Univ., 1982.

Flood, Robert P., and Marion, Nancy Peregrim. "The Transmission of Disturbances under Alternative Exchange-Rate Regimes with Optimal Indexing." *Q.J.E.* 97 (February 1982): 43–66.

Frenkel, Jacob A. "Flexible Exchange Rates, Prices, and the Role of 'News': Lessons from the 1970s." *J.P.E.* 89 (August 1981): 665–705.

Friedman, Milton. "The Role of Monetary Policy." *A.E.R.* 58 (March 1968): 1–17.

Harris, Richard G., and Purvis, Douglas D. "Diverse Information and Market Efficiency in a Monetary Model of the Exchange Rate." *Econ. J.* 91 (December 1981): 829–47.

Kimbrough, Kent P. "The Forward Rate as a Predictor of the Future Spot Rate, the Role of Policy, and Exchange Rate Regime Choice." Unpublished manuscript. Durham, N.C.: Duke Univ., 1982.

———. "The Information Content of the Exchange Rate and the Stability of Real Output under Alternative Exchange Rate Regimes." *J. Internat. Money and Finance* 2 (April 1983): 27–38. (*a*)

———. "Exchange Rate Policy and Monetary Information." *J. Internat. Money and Finance* 2 (December 1983): 333–46. (*b*)

King, Robert G. "Monetary Information and Monetary Neutrality." *J. Monetary Econ.* 7 (March 1981): 195–206.

———. "Monetary Policy and the Information Content of Prices." *J.P.E.* 90 (April 1982): 247–79.

———. "Interest Rates, Aggregate Information, and Monetary Policy." *J. Monetary Econ.* 12 (August 1983): 199–234.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24.

———. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34.

Lucas, Robert E., Jr., and Rapping, Leonard A. "Real Wages, Employment, and Inflation." *J.P.E.* 77 (September/October 1969): 721–54.

Mundell, Robert A. "Capital Mobility and Stabilization Policy und and Flexible Exchange Rates." *Canadian J. Econ. and Pol. Sci.* 29 (N 1963): 475–85.

Phelps, Edmund S. "The New Microeconomics in Employment and Theory." In *Microeconomic Foundations of Employment and Inflat* edited by Edmund S. Phelps et al. New York: Norton, 1970.

Sargent, Thomas J., and Wallace, Neil. "Rational Expectations, the Op Monetary Instrument, and the Optimal Money Supply Rule." *J.P.E.* (April 1975): 241–54.

Stockman, Alan C., and Koh, Ai Tee. "Models of Business Fluctuations in Open Economies." Unpublished manuscript. Rochester, N.Y.: Univ. Rochester, 1982.

Weiss, Laurence. "The Role for Active Monetary Policy in a Rational Expectations Model." *J.P.E.* 88 (April 1980): 221–33.

———. "Information Aggregation and Policy." *Rev. Econ. Studies* 49 (January 1982): 31–42.

International Capital Movements under Uncertainty

Gene M. Grossman

Princeton University

Assaf Razin

Tel-Aviv University

In this paper we analyze the determinants of international movements of physical capital in a model with uncertainty and international trade in goods and securities. In our model, the world allocation of capital is governed, to some extent, by the asset preferences of risk-averse consumer-investors. In a one-good variant in the spirit of the MacDougall model, we find that relative factor abundance, relative labor force size, and relative production riskiness have separate but interrelated influences on the direction of equilibrium capital movements. These same factors remain important in a two-good version with Heckscher-Ohlin production structure. In this case, the direction of physical capital flow is determinate (unlike in a world of certainty) and may hinge on the identity of the factor that is used intensively in the industry with random technology.

I. Introduction

The theory of international trade has been extended in recent years to incorporate uncertain trading environments. The early writers in

This paper was written while Grossman was visiting Tel-Aviv University. He thanks the National Science Foundation for its support under grant SES 8207643. Razin acknowledges the support of the Ross Endowment. We both are grateful to the Foerder Institute for Economic Research for financial assistance for the typing of this paper and to Avinash Dixit, Bill Ethier, Elhanan Helpman, Mike Katz, Victor Norman, José Scheinkman, Carl Shapiro, and Lars Svensson for comments on an earlier draft

[*Journal of Political Economy*, 1984, vol. 92, no. 2]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0003$01.50

this area (e.g., Kemp and Liviatan 1973; Turnovsky 1974; Batra 1975) argued that the introduction of randomness into the standard deterministic models had proven to be very damaging to many orthodox results, including those concerning the pattern of trade. However, as Helpman and Razin (1978*a*, 1978*b*) later showed, many of the negative findings were due to the implicitly assumed absence of markets for international risk sharing in those early models. When international trade in equities is admitted as a possibility, a number of the familiar theorems are restored. Under such conditions, if uncertainty takes the form of industry-specific (but not country-specific) multiplicative technological shift factors, and if certain restrictions are placed on agents' utility functions, then, as Anderson (1981) has shown, the usual comparative cost considerations reemerge as the determinants of the pattern of trade in securities.[1] Trade in commodities may also be predicted by the cost-based (i.e., Ricardian and Heckscher-Ohlin) theorems, at least in an "on average" sense.

International movements of physical capital, even more than the flow of goods, may be influenced by the existence of technological uncertainty. The worldwide allocation of capital takes place largely before the resolution of uncertainty and is motivated by, among other factors, the desire of risk-averse agents to hedge against risk. Capital flows not only into those sectors and countries where its expected marginal product is high, but also into those which singly or together provide investors with a relatively stable pattern of income across states of nature.

Risk considerations have long occupied a central position i
of the international flows of financial capital. The diversifica
tive seems equally relevant for movements of physical capi
indeed the hedging of uncertainty is often cited in informa
sions as the motive for instances of direct foreign investment
our knowledge, there has been no attempt in the trade literature
incorporate production risk explicitly into a formal model that determines the pattern of trade in (physical) capital.[2]

In this paper we study the interrelationship between international capital movements and international trade in securities under conditions of technological uncertainty. The problem takes on greatest interest under the assumption that random disturbances in an industry are not perfectly correlated across countries. Interestingly, the existence of such uncertainty introduces some fundamentally new elements into the determination of the direction and level of capital movements. Essentially, the general equilibrium supply functions for real equities derive from the familiar supply relationships, but the demands for equities also have an important qualitative effect on the equilibrium allocation of resources, even when all individuals in both countries have identical and homothetic tastes for goods and assets.

---

[1] Strictly speaking, when the random fluctuations in each industry are perfectly correlated across countries, comparative labor costs explain the pattern of trade in securities in a Ricardian-type model without any restrictions on utility functions (see Helpman and Razin 1978*a*, 1978*b*). The contribution of Anderson (1981) was to provide the conditions under which the Heckscher-Ohlin and Travis-Vanek theorems extend to situations of uncertainty

[2] Eaton and Gersovitz (1982) have developed a model of physical capital mobility incorporating the risk of expropriation. Their paper complements this one, which deals only with production risk.

We begin, in the next section, with a model in the spirit of MacDougall (1960). In additional to the usual influence of the autarky factor-endowment ratios, we find an important role for the relative sizes of the labor forces and for the distributions of the random technology variables in the determination of the volume and direction of capital movements. Indeed, capital may flow to the relatively capital-abundant country, even if the riskiness of the two countries as investment environments is the same.

In Section III, we extend the model to include an internationally traded bond. Under a restriction on the utility functions that is analogous to that needed to prove the Heckscher-Ohlin theorem (i.e., internationally identical and homothetic tastes for goods and assets), the introduction of the bond market does not alter the conclusions of Section II.

Finally, in Section IV, we investigate a two-good variant, adopting the framework of the Heckscher-Ohlin model. Whereas the nonrandom model is characterized by perfect substitution between commodity trade and factor movements (see Mundell 1957), and therefore by an indeterminacy in the level and direction of goods trade and capital movements, the equilibrium conditions under uncertainty determine nontrivially the volumes of all these flows (as well as trade in securities).[3] Under the assumption that technology in one of the industries in each country is nonstochastic, we are able to identify the separate roles of the relative sizes of the two labor forces and the relative factor intensities of the two industries in the determination of the direction of physical capital movements.

Our results are summarized in a concluding section.

---

[3] When industry-specific shocks are perfectly correlated across countries, factor-prices equalization obtains even in the cum-uncertainty model, provided that both countries are incompletely specialized in equilibrium (see Helpman and Razin 1978*a*). In this case, the intercountry allocation of the world's capital stock is again indeterminate.

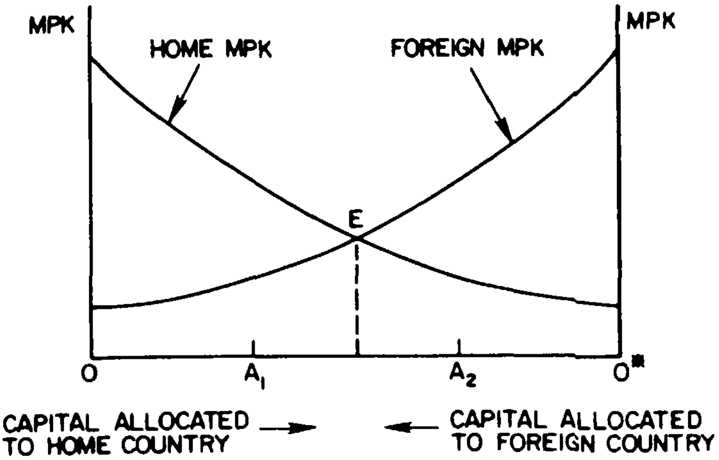

MPK
HOME MPK
FOREIGN MPK
MPK
E
O
A₁
A₂
O*
CAPITAL ALLOCATED TO HOME COUNTRY
CAPITAL ALLOCATED TO FOREIGN COUNTRY

FIG. 1

II. The MacDougall Model with Uncertainty

The simplest model in which capital movements can be analyzed is one with two countries, one good, and two factors. MacDougall (1960) developed such a model to study the welfare implications of capital movements in a deterministic world under a variety of assumptions about technology, the behavior of labor, market structure, and tax policy.

In this paper, we are interested only in the positive implications of the simplest variant of the MacDougall formulation (i.e., con returns-to-scale production functions, fixed labor supplies competition, and laissez-faire). In this case, equilibrium is c ized by equalization of the marginal products of capital in countries. Such an equilibrium is illustrated in figure 1, w horizontal dimension of the box represents the fixed work ment of capital, and the marginal product of capital as a fun the capital allocated to the home (foreign) country is plotted w respect to the origin at the left (right) of the figure. The equilibrium allocation is at E, while A_1 and A_2 are two possible autarky allocations. At A_1, the marginal product of capital at home exceeds that in the foreign country, whereas the opposite is true at A_2. Suppose production functions are the same in the two countries. Then A_1 must be characterized by a higher capital-to-labor ratio abroad than at home and A_2 by the reverse situation. Evidently, if technologies are the same and nonrandom, capital flows to the country that, in autarky, has a greater relative abundance of labor.

We begin our analysis of capital movements under uncertainty by introducing multiplicative technological randomness into each of the countries of the MacDougall model and allowing for international trade in equities in the manner of Helpman and Razin.

Let the S possible states of nature be indexed by α, $\alpha = 1, 2, \ldots, S$. Each state of nature is defined by the realization of two country-specific random variables, $\theta(\alpha)$ and $\theta^*(\alpha)$, corresponding to the state of technology in the home and foreign industries, respectively.[1] The output of the jth home-country firm in state α is $X_j(\alpha) = \theta(\alpha)F(L_j, K_j)$ for $\alpha = 1, 2, \ldots, S$ and $j = 1, 2, \ldots, J$, where F is a standard, quasi-concave, constant-returns-to-scale production function (the same for all firms), L_j is the firm's labor input, and K_j is its input of physical capital. Similarly the output of the jth foreign firm in state α is given by $X_j^*(\alpha) = \theta^*(\alpha)F(L_j^*, K_j^*)$ for $\alpha = 1, 2, \ldots, S$ and $j = 1, 2, \ldots, J^*$. The production function for foreign firms is identical to that for home firms, but the multiplicative uncertainty term, which is the same for all firms within each country, is not necessarily the same for firms located in different countries. With these assumptions, firms in each country can be aggregated to the industry level, so that henceforth we omit the j subscripts.

Firms in each country choose their inputs prior to the resolution of uncertainty. It is a well-known result that when technological uncertainty enters multiplicatively, as has been assumed here, stockholders are unanimous in their desire that firms act so as to maximize their net stock market values (i.e., gross values less factor payments). Let q and q^* be the prices of a unit of real equity in a representative home and foreign firm, respectively, with $q \equiv 1$ by choice of numeraire. A unit of real equity in a home firm pays $\theta(\alpha)$ units of the (single) consumption good if state α is realized. Similarly, $\theta^*(\alpha)$ is the return to a unit of the foreign equity. Home and foreign firms produce $Z = F(L, K)$ and $Z^* = F(L^*, K^*)$ units of real equities, respectively, which have gross stock market values of $F(L, K)$ and $q^*F(L^*, K^*)$. Thus, the home-country industry chooses L and K to maximize $F(L, K) - wL - rK$, where w is the home wage rate and r the home rental rate for capital, both expressed in terms of home equities. The first-order conditions for maximization are

$$F_L(L, K) = w, \tag{1}$$

$$F_K(L, K) = r. \tag{2}$$

The foreign industry seeks to maximize $q^*F(L^*, K^*) - w^*L^* - r^*K^*$ and thus chooses L^* and K^* to satisfy

---

[1] We use asterisks to refer to variables for the foreign country

$$q^*F_L(L^*, K^*) = w^*, \tag{3}$$

$$q^*F_K(L^*, K^*) = r^*. \tag{4}$$

Capital and labor endowments in the home and foreign countries are $\bar{K}$ and $\bar{K}^*$, respectively, and labor endowments are $\bar{L}$ and $\bar{L}^*$. Labor is internationally immobile, so that the labor markets must clear separately in each country. In equilibrium, we have

$$L = \bar{L}, \tag{5}$$

$$L^* = \bar{L}^*. \tag{6}$$

Capital movements are costless and unrestricted, which implies the existence of a unified, world, physical capital market. The conditions for equilibrium in this market are

$$K + K^* = \bar{K} + \bar{K}^*, \tag{7}$$

$$r = r^*. \tag{8}$$

We turn finally to consumer behavior. Consumer-investors in each country are endowed with physical capital, labor, and shares of ownership in firms. Prior to the resolution of uncertainty, each individual sells his factor endowments, bears his fraction of each firm's factor costs in accordance with his initial ownership, and buys and sells shares of stock in the various firms.

Let $V^i[I^i(\alpha)]$ be the concave, von Neumann–Morgenstern (indirect) utility function for individual i, where $I^i(\alpha)$ is the individual's income in state α. Suppose the individual were to hold in his ultimate portfolio z^i shares of stock in home firms and z^{*i} shares in foreign firms. Then his income in state α would be $I^i(\alpha) = \theta(\alpha)z^i + \theta^*(\alpha)z$ portfolio choice problem of this individual is to maximize ex utility, given (common) subjective beliefs about the probabili bution for the states of nature and subject to the budget cons that the cost of his portfolio not exceed the value of his initial ment. That is, the individual solves $\max_{z^i, z^{*i}} EV^i[\theta(\alpha)z^i + \theta^*(\alpha)$ ject to $z^i + q^*z^{*i} = wL^i + rK^i$, where L^i and K^i are the individu initial endowments of labor and capital, respectively. Note that with constant returns to scale, the equilibrium value of any initial shareholdings by this individual is, after factor payments are made, equal to zero. The portfolio allocation that solves this problem (assuming an interior solution) must satisfy

$$\frac{E\theta^*(\alpha)V_I^i[\theta(\alpha)z^i + \theta^*(\alpha)z^{*i}]}{E\theta(\alpha)V_I^i[\theta(\alpha)z^i + \theta^*(\alpha)z^{*i}]} = q^*, \tag{9}$$

where $V_I^i(\cdot)$ is the marginal utility of income.

The model is closed by the world market-clearing conditions for the real equities of firms located in each of the countries; that is,

$$\sum_i z^i = Z, \tag{10}$$

$$\sum_i z^{*i} = Z^*, \tag{11}$$

where the summation is over all individuals in the world.

Before proceeding to an investigation of the properties of the cum-factor-movements equilibrium, we choose to place a restriction on the form of the utility functions that is analogous to the one often invoked in nonstochastic trade models for proofs of theorems on the determinants of commodity trade. In the present context, the assumption is that all consumers, worldwide, have identical and homothetic preferences over equities. The purpose of this assumption is to neutralize any bias in the pattern of trade in securities or in the direction of capital movements introduced on the demand side by differences in tastes or by income distributional considerations.[5]

For consumers' preferences over securities to be identical and homothetic, it is sufficient that they all have utility functions that exhibit identical and constant relative aversion to income risk; that is, that their utility functions be of the form $V(\cdot) = (I^i)^{1-\gamma}/(1 - \gamma)$, for some $\gamma \neq 1$, or of the form $V(\cdot) = \log (I^i)$. If the utility function takes one of these forms, (9) can be rewritten as

$$\frac{E\theta^*(\alpha)V_I[\theta(\alpha) + \theta^*(\alpha)\tilde{z}^i]}{E\theta(\alpha)V_I[\theta(\alpha) + \theta^*(\alpha)\tilde{z}^i]} = q^* \tag{9$'$}$$

(where $\tilde{z}^i = z^{*i}/z^i$), from which it is clear that the relative holdings of the two stocks in any investor's portfolio are independent of his nationality or level of wealth.

Identical, homothetic preferences have the property of being aggregable. That is, world demand for assets can be consistently represented by a set of community asset indifference curves of the form $EV[\theta(\alpha)z + \theta^*(\alpha)z^*] = \bar{V}$. These also represent demand in each country taken separately. Utility is a quasi-concave function of asset holdings, and is strictly so, if individuals are risk averse (i.e., $V_{II} < 0$) and if $\theta(\alpha)$ and $\theta^*(\alpha)$ are less than perfectly correlated.

The nature of the world equilibrium with free capital movements is

---

[5] Anderson (1981) was the first to recognize the relevance of this assumption in the context of the Helpman-Razin model of trade in goods and securities. When, in Sec IV, we introduce a second good, our assumption will be that preferences over goods are identical and homothetic as well.

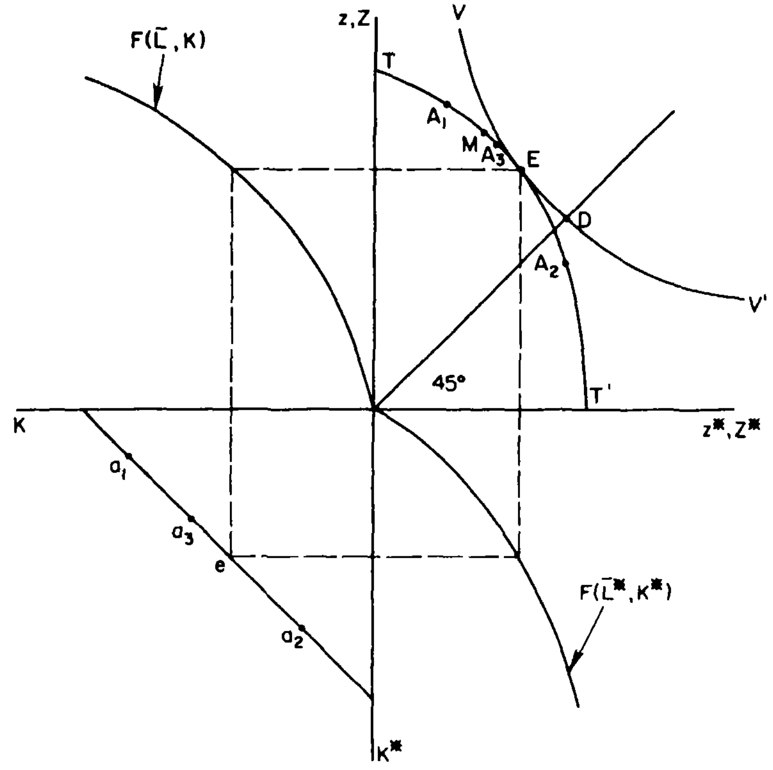


FIG. 2

best understood with the aid of figure 2. For illustrative pu
depict a situation in which $\theta(\alpha)$ and $\theta^*(\alpha)$ have identical dist
In quadrants II and IV, we draw the production function
equities in each country as a function of the amount of capita
there. The sum of the allocations of capital to the two countries constrained by equation (7) to be equal to the exogenous world supply. This constraint is represented by a straight line with a slope of -1 in quadrant III. Together, these constraints trace out, in quadrant IV, a world transformation locus, TT', relating the feasible supplies of the two real equities. Each point on the frontier corresponds to a particular division of capital between the two countries. The slope of the transformation curve at any point is given by $-F_K(\bar{L}, K)/F_K(\bar{L}^*, K^*)$.

---

[6] In other words, we suppose for the sake of the diagram that the joint density function for the two random variables, $\psi(\cdot, \cdot)$, satisfies $\psi(\theta, \theta^*) = \psi(\theta^*, \theta)$

A representative of the family of homothetic asset indifference curves is depicted in quadrant IV by VV'. The slope of VV' is given by (the negative of) the marginal rate of substitution of assets, that is, by the left-hand side of equation (9). Under the assumption that $\theta(\alpha)$ and $\theta^*(\alpha)$ are symmetrically distributed, the slope of VV' must be -1 where $z = z^*$.

World equilibrium occurs at E, the point of tangency of an asset indifference curve and the world transformation locus. This is because consumers set (the negative of) the marginal rate of substitution between assets (MRSA) equal to the relative price of securities (eq. [9]), while competition in the world market for physical capital leads to equality between (the negative of) the marginal rate of transformation (MRTA) and the relative price of securities (eqq. [2], [4], [8]).

In the diagram, the deterministic equilibrium, or "MacDougall point," is labeled M. As we have noted, the deterministic equilibrium is characterized by $F_K(\bar{L}, K) = F_K(\bar{L}^*, K^*)$, so the marginal rate of transformation at M is -1. By the concavity of the asset indifference curve and the fact that the latter has a slope of -1 at point D, it follows that equilibrium with uncertainty must lie (weakly) between the MacDougall point and the 45° line; that is, it must exhibit more diversification. If agents are risk neutral, or if $\theta(\alpha)$ and $\theta^*(\alpha)$ are perfectly correlated, then the asset indifference curves are straight lines with slopes of -1, and the two equilibria coincide at M. Alternatively, if $\bar{L} = \bar{L}^*$, the MRTA is -1 at the 45° line, and again the equilibria coincide. In all other cases, the equilibrium under uncertainty differs from that for the case of no uncertainty by an amount that depends on the degree of relative risk aversion and the correlation between disturbances in the two countries. A higher degree of relative risk aversion and a less positive (or more negative) correlation between $\theta(\alpha)$ and $\theta^*(\alpha)$ tend to render the asset indifference curves more concave and thus contribute to a larger distance between points M and E in the figure.

The introduction of uncertainty may alter the nature of the capital-movements equilibrium either quantitatively or qualitatively. Consider the three potential autarky points at A_1, A_2, and A_3. (Autarky production is represented by the point along TT' that corresponds to the exogenously given initial endowment at, e.g., a_1, a_2, or a_3.) If autarky production is at A_1, then in both the deterministic model and the model with uncertainty the home country imports capital, but *more* capital movement takes place in the latter case. With autarky at A_2, capital flows out of the home country in both situations, but now the introduction of uncertainty *lessens* the extent of capital movement. Finally, if autarky is at A_3, the introduction of uncertainty reverses the direction of capital movement relative to the outcome in a deterministic world.

What can be said in general about the direction of capital movement in a one-good world with uncertainty and international trade in securities? In order to isolate the separate influences of the relative size of the labor forces, the relative factor abundances, and the relative riskiness of the two countries, we consider, in turn, initial situations that deviate from complete symmetry along only one of these dimensions. Our findings are summarized in three propositions.

PROPOSITION 1: If $\theta(\alpha)$ and $\theta^*(\alpha)$ have identical distributions and $\bar{L} = \bar{L}^*$, then physical capital moves toward the country with the smaller autarky endowment of capital.

PROOF: The fact that $\theta(\alpha)$ and $\theta^*(\alpha)$ have identical distributions implies that the MRSA is -1 for $z = z^*$. The equality of labor forces implies that the MRTA is -1 for $Z = Z^*$. Thus, an equal division of the world's capital stock across countries ($K = K^*$) satisfies all the conditions for equilibrium. Also, equilibrium is unique, so capital must flow from the country in which it is initially abundant toward the less well endowed country.

PROPOSITION 2: If $\theta(\alpha)$ and $\theta^*(\alpha)$ have identical distributions and $\bar{K}/\bar{L} = \bar{K}^*/\bar{L}^*$, then physical capital moves toward the country with the smaller labor force.

PROOF: If $\bar{L}$ is greater (less) than $\bar{L}^*$, the autarky point lies along TT' above (below) the point where $Z = Z^*$. The MRTA is -1 at the autarky point—recall that MRTA $= -F_K(\bar{L}, K)/F_K(\bar{L}^*, K^*)$ and that F is homogeneous of degree one—and decreases monotonically for movements downward along TT'. The slope of the asset indifference curve that intersects TT' is -1 at the point where $Z = Z^*$ and increases monotonically for movements downward along the transformation locus. It follows that equilibrium must lie on TT' betw
autarky point and the point where $Z = Z^*$. In equilibrium
more production of the real equity of the initially smaller
than there is in autarky; that is, capital moves toward the cou
the smaller labor force.

PROPOSITION 3: If $\bar{K} = \bar{K}^*$ and $\bar{L}$ and $\bar{L}^*$, and the distribut
random variable in the foreign country is riskier than the distri
in the home country—in the sense that $\theta^*(\alpha) = \tilde{\theta}^*(\alpha) + \epsilon(\alpha)$, where $\theta(\alpha)$ and $\tilde{\theta}^*(\alpha)$ have identical distributions, $\epsilon(\alpha)$ is independent of both $\theta(\alpha)$ and $\tilde{\theta}^*(\alpha)$, and $E[\epsilon(\alpha)] = 0$[7]—then capital moves toward the home country.

---

[7] Note that this definition of "riskier" is more restrictive than the notion of a mean-preserving spread, as developed by Rothschild and Stiglitz (1970). A mean-preserving spread of $\tilde{\theta}^*(\alpha)$ would require only $E[\epsilon(\alpha) \mid \tilde{\theta}^*(\alpha)] = 0$. We have imposed the additional restriction that $\epsilon(\alpha)$ and $\tilde{\theta}^*(\alpha)$ be independent. We also limit our attention to cases where the noise that is added to $\tilde{\theta}^*(\alpha)$ is independent of $\theta(\alpha)$. Thus, our definition of riskier implies, but is not implied by, a mean-preserving spread.

Proof: *The asset indifference curve that intersects TT' at the point where $Z = Z^*$ has a slope given in (9′) by*

$$-\frac{E[\tilde{\theta}^*(\alpha) + \epsilon(\alpha)]V_I[\theta(\alpha) + \tilde{\theta}^*(\alpha) + \epsilon(\alpha)]}{E\theta(\alpha)V_I[\theta(\alpha) + \tilde{\theta}^*(\alpha) + \epsilon(\alpha)]} = -\left\{1 + \frac{E\epsilon(\alpha)V_I[\theta(\alpha) + \tilde{\theta}^*(\alpha) + \epsilon(\alpha)]}{E\theta(\alpha)V_I[\theta(\alpha) + \tilde{\theta}^*(\alpha) + \epsilon(\alpha)]}\right\},$$

where this equality follows from the fact that $\theta(\alpha)$ and $\tilde{\theta}^*(\alpha)$ are identically distributed and each is independent of $\epsilon(\alpha)$. The second term in the brackets is negative, since $\epsilon(\alpha)$ and $V_I(\cdot)$ have negative covariance. Thus, the asset indifference curve has a slope greater than -1 at $z = z^*$, while the slope of TT' is equal to -1 there. It follows that equilibrium is at a point above the 45° line, that is, that capital moves to the less risky country.

In general situations, the direction of capital movement is determined by the interaction of the separate influences of relative country size, relative factor abundance, and relative country riskiness. However, the nature of this interaction can be quite complex. It is not true, for example, that an increase in the riskiness of one country (in a mean-preserving spread sense) will always cause capital to flow out of that country. If the country that becomes riskier also has a smaller labor force and if the disturbances in the two countries are negatively correlated, then the increase in riskiness makes the real equity of that country a more attractive asset. Intuitively, the extra income the asset provides when the marginal utility of income is high outweighs the utility cost of the income forgone when marginal utility is low. One general statement that can be made is that, ceteris paribus, more capital will flow into a country the smaller is its labor force. The desire for diversification on the part of consumer-investors implies a tendency for real equity supplies to be equalized.

III. The MacDougall Model with a Traded Bond

In the previous section we studied capital movements in a one-good model with uncertainty, where the only assets available to consumer-investors were risky real equities. In the present section, we extend the analysis to incorporate a market for a safe asset (i.e., an internationally traded bond), while maintaining all of our earlier assumptions, including especially the one restricting agents' asset preferences to be identical and homothetic. We will show that this extension does not alter any of the conclusions of the previous section.

Let b^i be the holding of an internationally traded bond, with price, q_b, by the ith individual. This asset pays a return of one unit of the consumption good in all states of nature. The consumer-investor must allocate his ex ante wealth over three assets, the two real equities, and the bond. The first-order conditions for expected utility maximization imply, in place of equation (9), the following equations:

$$\frac{E\{\theta^*(\alpha)V_i[\theta(\alpha)z^i + \theta^*(\alpha)z^{*i} + b^i]\}}{E\{\theta(\alpha)V_i[\theta(\alpha)z^i + \theta^*(\alpha)z^{*i} + b^i]\}} = q^*, \tag{12}$$

$$\frac{E\{V_i[\theta(\alpha)z^i + \theta^*(\alpha)z^{*i} + b^i]\}}{E\{\theta(\alpha)V_i[\theta(\alpha)z^i + \theta^*(\alpha)z^{*i} + b^i]\}} = q_b. \tag{13}$$

The bond-market-clearing condition is

$$\sum_i b^i = 0. \tag{14}$$

All of the remaining equilibrium conditions of the earlier setup continue to apply.

The fact that all individuals have identical and homothetic demands for assets implies that, in equilibrium, each will allocate the same fraction of his wealth to any given asset. If the equilibrium bond holding of one individual is either strictly positive or strictly negative, such would also be true for every other individual. In either case, equation (14) could not be satisfied. It follows that, in equilibrium, $b^i = 0$ for all i. Once this fact is recognized, it is clear that the asset holdings that satisfy equation (9) will also satisfy equation (12). Indeed, all the conditions of the capital-movements equilibrium in the absence of bond trading are also consistent with equilibrium [illegible] bond market is assumed to exist. Equation (13), then, serve[illegible] mine the price of bonds such that in equilibrium all agents [illegible] take a net position of zero in the market for this safe asset [illegible] marize this finding in the following proposition.

PROPOSITION 4: When all agents have identical, homoth[illegible] utility functions, the equilibrium allocation of resources with fre[illegible] ital movements and free trade in equities and an "inside" bond is identical to the equilibrium allocation when the bond market does not exist.

IV. The Heckscher-Ohlin-Mundell Model with Uncertainty

In a deterministic world with two goods and two factors, international trade in goods and international factor movements are perfect substitutes, provided that both countries are incompletely specialized in the ultimate equilibrium (see Mundell 1957). In other words, the factor price equalization theorem implies that if capital movements are allowed, starting from an equilibrium with free trade in goods, no movements are actually needed to maintain equilibrium.[8]

In the present section we extend the model of Section II to incorporate a second consumption good, and therefore a second security, in each country. The resulting model is a cum-uncertainty analogue of the deterministic Heckscher-Ohlin model with free capital movements. In contrast to the results of Mundell for the nonrandom case, we find that goods trade is not a perfect substitute for factor movements and that equilibrium generally involves transactions in both international markets in determinate amounts. The reason for this difference is that without uncertainty the location of production of a good is immaterial; with uncertainty the location of production is economically relevant when equities are imperfect substitutes.

We assume that two goods are produced in each country with capital and labor and adopt completely the Heckscher-Ohlin production structure with regard to equity outputs. As before, the outputs of one of the goods (good 1) are stochastic and are given by $X = \theta(\alpha)F(L_x, K_x)$ in the home country and $X^* = \theta^*(\alpha)F(L_x^*, K_x^*)$ in the foreign country. The outputs of real equities in this industry are once again denoted Z and Z^*, with equity prices q_x ($\equiv 1$) and q_x^*. For simplicity, we assume that the technology for the second good is nonstochastic. Then the outputs of real equities in this industry are identically equal to the outputs of good 2 and are related to factor inputs by $Y = G(L_y, K_y)$ and $Y^* = G(L_y^*, K_y^*)$. Shares of stock in firms in industry 2 are perfect substitutes, irrespective of the country of location. Let q_y denote the common equity price for shares of firms in this industry, measured relative to the price of the home-country real equities of industry 1. Finally, let $p(\alpha)$ be the relative price of good 1 in terms of good 2 in state of nature α.

As is known from the works of Helpman and Razin, the conditions for production equilibrium are the same as those in the standard Heckscher-Ohlin model, except that real equity prices substitute here for commodity prices. Of course, with capital internationally mobile, there is the additional condition that the rental rates for capital be equalized across countries. As in the deterministic model, an equilibrium with incomplete specialization in each country will turn out to be possible so long as the ultimate factor endowment ratios are not "too" disparate. Since capital is freely mobile, this implies only a condition on the comparative sizes of the two labor forces. We will concentrate our attention on those situations where an equilibrium with incomplete specialization is possible.

---

[8] Similarly, goods trade need not emerge if the initial situation is one of free capital movements. As Mundell (1957) has shown, this implies that any tariff on goods will be prohibitive when capital is internationally mobile, provided that a situation of incomplete specialization is consistent with equilibrium.

Consider now the optimization problem of the typical consumer-investor. This individual wishes to maximize the mathematical expectation of his ex post utility, defined over his levels of consumption of the two goods, $c_x^i(\alpha)$ and $c_y^i(\alpha)$. Let $U[c_x^i(\alpha), c_y^i(\alpha)]$ represent his ex post utility function. Anderson (1981) has shown that, to ensure that all individuals have identical homothetic commodity *and* equity preferences, it is necessary and sufficient to restrict the form of $U(\cdot, \cdot)$ to be any linear transform of a positive power function (of the same degree for all individuals) of a homogeneous, quasi-concave function (also identical for all individuals).

The individual's maximization can be thought of as consisting of two stages. After the uncertainty is resolved, the individual qua consumer who holds a portfolio of z^i shares of firms in the home industry 1, z^{*i} shares of firms in the foreign industry 1, and y^i shares of firms in industry 2 in either country has income $I^i(\alpha) = \theta(\alpha)p(\alpha)z^i + \theta^*(\alpha)p(\alpha)z^{*i} + y^i$.[9] He allocates this income to maximize ex post utility according to the familiar condition

$$\frac{U_{c_x}[c_x^i(\alpha), c_y^i(\alpha)]}{U_{c_y}[c_x^i(\alpha), c_y^i(\alpha)]} = p(\alpha).$$

The solution to this second-stage problem can be used to define an indirect utility function, $V[p(\alpha), I^i(\alpha)]$.

In the first stage (i.e., prior to the resolution of uncertainty), the individual qua investor sells his initial endowment and allo
proceeds among the three assets to maximize $EV[p(\alpha), I^i(\alpha$
sume that price expectations are formed rationally at this st
$\equiv z^{*i}/z^i$ and $\tilde{y}^i \equiv y^i/z^i$. Then the first-order conditions for
utility maximization imply

$$\frac{E\{\theta^*(\alpha)p(\alpha)V_I[p(\alpha), \theta(\alpha)p(\alpha) + \theta^*(\alpha)p(\alpha)\tilde{z} + \tilde{y}]\}}{E\{\theta(\alpha)p(\alpha)V_I[p(\alpha), \theta(\alpha)p(\alpha) + \theta^*(\alpha)p(\alpha)\tilde{z} + \tilde{y}]\}} = q_z$$

and

$$\frac{E\{V_I[p(\alpha), \theta(\alpha)p(\alpha) + \theta^*(\alpha)p(\alpha)\tilde{z} + \tilde{y}]\}}{E\{\theta(\alpha)p(\alpha)V_I[p(\alpha), \theta(\alpha)p(\alpha) + \theta^*(\alpha)p(\alpha)\tilde{z} + \tilde{y}]\}} = q_y, \qquad (16)$$

where the fact that asset preferences are identical and homothetic allows us to drop the i superscripts.

---

[9] Note that the return to a unit of real equity is $\theta(\alpha)p(\alpha)$ for a firm in the home industry 1 and is $\theta^*(\alpha)p(\alpha)$ for a firm in the foreign industry 1. A unit of real equity for a firm in industry 2, irrespective of location, pays a return of one unit of the numeraire good.

The model is closed, as before, by market-clearing conditions for assets (e.g., $\Sigma_i y^i = Y + Y^*$, $\Sigma_i z_i = Z$, etc.). These are augmented now by ex post market-clearing conditions state by state for each of the goods, that is, $\Sigma_i c^i_x(\alpha) = X(\alpha) + X^*(\alpha)$ and $\Sigma_i c^i_y(\alpha) = Y + Y^*$.

We will henceforth restrict our attention to situations where the distributions of the country-specific random variables are identical (though not perfectly correlated). To understand the qualitative nature of the equilibrium in such cases, consider first the ex post relative commodity price that is realized in state of nature α. By the homotheticity of aggregate demands, this market-clearing price is a function only of the relative outputs of the two goods, that is, $p(\alpha) = \pi\{[\theta(\alpha)Z + \theta^*(\alpha)Z^*]/(Y + Y^*)\}$. Also, in equilibrium $\bar{z} = Z^*/Z$ and $\bar{y} = (Y + Y^*)/Z$. Thus

$$p(\alpha) = \pi\left\{\frac{[\theta(\alpha) + \theta^*(\alpha)\bar{z}]}{\bar{y}}\right\}. \tag{17}$$

Substituting (17) into (15), we note that if the prices of the two industry 1 assets are the same (i.e., $q^*_x = 1$), then in equilibrium equal quantities of the two assets will be held by every consumer-investor (i.e., $\bar{z} = 1$). This is because when $Z = Z^*$, $\pi(\cdot)$ is symmetric in $\theta(\alpha)$ and $\theta^*(\alpha)$, as is $V_i(\cdot)$. Since by assumption $\theta(\alpha)$ and $\theta^*(\alpha)$ have identical distributions, the marginal rate of asset substitution is one when the securities are demanded equally. Furthermore, by the quasi concavity of the asset indifference curves, $\bar{z} > 1$ (< 1) in equilibrium if and only if $q^*_x < 1$ (> 1). Investors demand more of the asset that has the lower price.

As we shall show presently, two qualitatively different equilibrium configurations are possible. If the two countries are of relatively equal size (as measured by their labor forces), then equilibrium will entail incomplete specialization in both countries. Alternatively, if one country is much larger than the other, the smaller country will specialize its production in the risky industry (see fig. 3).

Consider a potential equilibrium in which neither country spe-

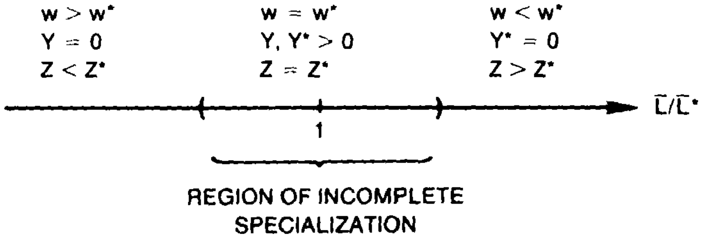


Fig. 3

cializes in the risky industry. In such circumstances, the countries produce an identical asset in common (the industry 2 security). Since perfect capital mobility implies $r = r^*$ always, it follows that full factor price equalization (with $w = w^*$) must obtain in this case. Then, since (equity) price equals unit cost for the good 1 industry in both countries[10] and factor prices are equalized, we must have $q_x^* = q_x$ ($\equiv 1$). Finally, recalling our discussion of asset demands, we conclude that an equilibrium with incomplete specialization in both countries is characterized by $Z = Z^*$. Note that such an outcome must result, for example, when $\bar{L} = \bar{L}^*$, in which case an equal distribution of the world's capital stock across countries and symmetric production patterns satisfies all the conditions for equilibrium.

The equalization of absolute levels of the output of the real equity of industry 1 is a key property of an equilibrium with incomplete specialization, and one that we shall exploit to advantage in our discussion of the determinants of the direction of international capital movements. There is, of course, no such requirement for outputs of good 1 in the two countries of a deterministic model. The difference arises because, whereas the outputs of the first industry (i.e., good 1) are perfect substitutes in a nonstochastic world, they no longer are so when what firms produce are real equities in a world of uncertainty. As will become evident in the course of our discussion, this difference explains why factor movements do not of necessity occur under certainty, given free trade in goods, but do generally occur with uncertainty.

Suppose now that one country (e.g., the foreign country) specialized its production in the risky industry in equilibrium. Then the fact that $Y > 0$ and $Y^* = 0$ implies, by the supply conditions for equity 2, that $w^* \geq w$. Then, from price equals unit cost in industry 1, we have $q_x^* \geq 1$. This in turn implies $Z > Z^*$ by consideration of asset dem[illegible] Finally, since $L_x/Z \geq L_x^*/Z^*$ (recall relative factor prices in [illegible] countries), it must be the case that $\bar{L} > \bar{L}^*$. By similar reason[illegible] equilibrium with $Y^* > 0$ and $Y = 0$ can occur only when $\bar{L}^*$ [illegible] an equilibrium would have $w^* \leq w$ and $q_x^* \leq 1$.

This completes our discussion of the qualitative nature of t[illegible]ous possible equilibrium configurations. It remains only to inves[illegible] the patterns of capital movements and international trade that are implied by these equilibria. These patterns depend, of course, on the initial distribution of the world's physical capital. As before, let us proceed by considering in turn initial situations that deviate from perfect symmetry between countries along only one dimension. First, we have:

---

[10] We must also consider the possibility that one country (say the foreign country) specializes in the safe industry. Then the security price for the foreign industry 1 must (weakly) exceed unit cost, which implies $q_x^* \geq 1$. But then $Z > Z^* = 0$ is inconsistent with equilibrium in the asset markets, since the foreign equity has both lower price and smaller quantity.

Proposition 5: If $\theta(\alpha)$ and $\theta^*(\alpha)$ are identically distributed and $\bar{L} = \bar{L}^*$, capital moves toward the country with the smaller endowment of capital. That country has positive expected exports of both goods. The ratio of its expected exports is equal to the ratio of its expected consumption levels.

Proof: With $\bar{L} = \bar{L}^*$, the equilibrium conditions $Z = Z^*$ and $q_x^* = 1$ (which hold for an incomplete-specialization equilibrium) are satisfied if $K_x + K_y = K_x^* + K_y^*$, that is, if the world's capital stock is equally allocated across countries. Thus, capital moves toward the country that has the smaller endowment.

Consider now the pattern of ex post goods trade. Let $C_h(\alpha)$ and $C_h^*(\alpha)$ be the aggregate consumption levels of good h ($h = X, Y$) in the home and foreign countries, in state α. For definiteness, assume that $\bar{K} > \bar{K}^*$ and define $\lambda \equiv [(w\bar{L} + r\bar{K})/(w\bar{L}^* + r\bar{K}^*)] - 1$. The percentage excess of home endowment wealth over foreign endowment wealth is λ.[11] The homotheticity of commodity preferences implies that $C_h(\alpha) = (1 + \lambda)C_h^*(\alpha)$. Since ex post factor allocations are equal, $EX(\alpha) = EX^*(\alpha) = [EC_x(\alpha) + EC_x^*(\alpha)]/2$ and $Y = Y^* = [EC_y(\alpha) + EC_y^*(\alpha)]/2$. Expected exports of good 1 by the foreign country are

$$EX^*(\alpha) - EC_x^*(\alpha) = \frac{\lambda EC_x^*(\alpha)}{2} > 0.$$

Its expected exports of good 2 are

$$Y^* - EC_y^*(\alpha) = \frac{\lambda EC_y^*(\alpha)}{2} > 0.$$

The surplus on trade account is balanced by a deficit on service account.

Rearrangement of the ex post market-clearing condition for good 2 gives

$$Y^* - C_y^*(\alpha) = \frac{(1 + \lambda)Y^* - Y}{2 + \lambda}.$$

Thus, the volume of trade in good 2 is state independent, whereas the flow of good 1 varies across states of nature. We note, however, that the *expected* trade pattern can be described as "neutral" and clearly is not determined by initial relative factor abundance. This is similar to the result reported by Svensson (forthcoming), who found in a different context that trade in goods might not be explainable by autarky relative factor abundance when factors are internationally mobile.

---

[11] Note that the existence of free capital movement and free trade in equities implies, in an equilibrium with incomplete specialization, that wage rates are equalized. Also, in equilibrium the value of initial stock holdings is zero (see Helpman and Razin 1978*a*, chap. 4).

PROPOSITION 6: Suppose $\theta(\alpha)$ and $\theta^*(\alpha)$ are identically distributed and $\bar{K}/\bar{L} = \bar{K}^*/\bar{L}^*$. Capital moves toward the smaller country *if* the industry with random technology is relatively capital intensive and toward the larger country otherwise. The smaller country always imports good 2; if it is incompletely specialized in equilibrium, it has positive expected exports of good 1.

PROOF: The proof makes use of figure 4, where we have assumed, for definiteness, that the home country is the larger one. In each panel of the figure we depict the real-equity production possibility frontiers (EPPF) for the two countries under autarky. Since relative factor endowments are equal, the EPPF for the larger country is a radial expansion of the EPPF for the smaller country.

Suppose, to begin with, that an equilibrium with incomplete specialization is possible. Then, in equilibrium, $Z = Z^*$. Consider the hypothetical real-equity production that would take place in each country at the relative equity prices that ultimately prevail in the equilibrium with free capital movements if capital movements were zero. These points are marked B and B^* in each panel of the figure. Clearly, the absence of capital movements is inconsistent with equilibrium, since at these points $Z > Z^*$.

Now suppose that the industry with uncertain technology
tively capital intensive (see fig. 4, *a*). Imagine a transfer of cap
the home to the foreign country, holding the relative equi
constant at their equilibrium values. This would cause the tw

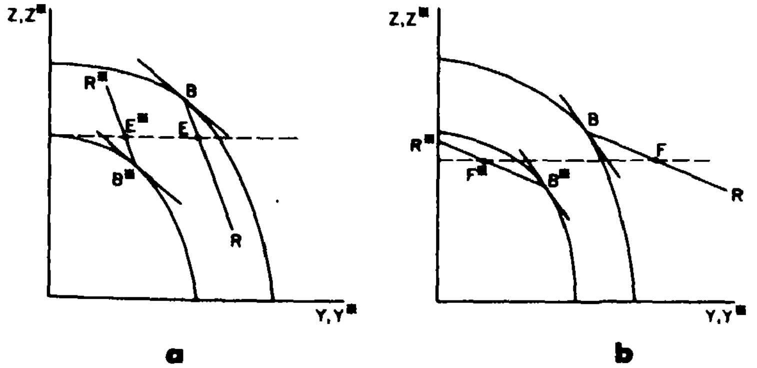


FIG. 4

tries to move in opposite directions along their respective capital-Rybczynski lines, *BR* and B^*R^*.[12] That is, the reallocation of capital causes the foreign country to expand its production of the capital-intensive security (that of industry 1, in this case) and to contract its production of the labor-intensive security, while the home country does just the opposite. The Rybczynski lines are parallel, because the equality of relative equity prices in equilibrium implies that factor prices are equal for this hypothetical transfer. Eventually points *E* and E^* are reached, such that $Z = Z^*$. These real-equity production points are consistent with full equilibrium.

In figure 4, *b*, the industry with uncertain technology is assumed to be labor intensive. Then, equalization across countries of the real-equity outputs of industry 1 requires a transfer of capital from the small country to the large country. The equilibrium production points are at *F* and F^*. It is perhaps counterintuitive that capital moves in this case toward the larger country, even though factor endowment ratios are equal and country risks are symmetric. This direction of capital movement is dictated by the strong implications of the Rybczynski theorem for the output effects of endowment changes.

If the sizes of the two countries (measured by labor force) are disparate enough, an equilibrium with incomplete specialization will not be possible. In terms of the diagram, the small country, in moving along its Rybczynski line, will hit the vertical axis before a point is reached where $Z = Z^*$. In such instances the small country will specialize in production of good 1 and the large country may or may not specialize in good 2. Nonetheless, the direction of capital movement must be as described in the proposition.

Turning to the pattern of commodity trade, the foreign country has exports of good 2 given by

$$Y^* - C_Y^* = \frac{(1 + \lambda)Y^* - Y}{2 + \lambda}.$$

Referring once again to figure 4, note that in each panel the output of the equity of industry 2 at point *B* is $1 + \lambda$ times as large as the output of this equity at point B^*. In both panels capital flows in the direction that causes home output of good 2 to increase and foreign output to contract. Thus, in any equilibrium with incomplete specialization (as depicted), $Y > (1 + \lambda)Y^*$, and the smaller country imports good 2 in every state of nature. Obviously, the smaller country also imports good 2 in cases where that country specializes in the risky industry.

---

[12] Helpman and Razin (1978*a*, 1978*b*) have shown that the Rybczynski theorem is applicable to real equity outputs when securities are internationally traded.

In the event of incomplete specialization, expected exports by the foreign country of the product of the risky industry are

$$EX^*(\alpha) - EC_x^*(\alpha) = \frac{\lambda C_x^*(\alpha)}{2} > 0,$$

where we have made use of the fact that $EX(\alpha) = EX^*(\alpha)$ in such an equilibrium. By contrast, when the smaller country is completely specialized, we have $EX^*(\alpha) < EX(\alpha)$, and it might be the case that this country has positive (expected) imports of both goods.

Note that once again the pattern of (expected) commodity trade cannot be predicted on the basis of initial relative factor abundance.

V. Conclusions

When production is characterized by technological uncertainty that is, at least to some extent, country specific, the international allocation of mobile factors is influenced by the asset preferences of risk-averse consumer-investors. In this paper we have studied the determinants of the direction of international capital movements in a model of trade in commodities and real equities under the assumption that preferences over commodities and assets are identical and homothetic worldwide. In a one-good variant of our model in the spirit of MacDougall (1960), we found that physical capital flows in an uncertain world are subject to the combined influences of relative factor abundance, relative size of labor force, and relative country riskiness. When deviation from complete symmetry is along only one of these dimensions, capital moves toward the relatively labor-abundan
try, the smaller country, and the less risky country, respectivel
ever, in more general situations, the interaction between these
can be quite complex.

Many of the lessons from the MacDougall model remain ap
when the model is extended to incorporate a second good
produced in each country with a nonstochastic technology. I
trast to the deterministic two-good, two-factor model with poten
capital mobility, equilibrium under uncertainty generally requires some movement of capital even when goods are traded freely. Relative factor abundance, relative size of labor force, and relative country riskiness still play important roles in the determination of the direction of factor movements, as well as in the determination of the expected pattern of commodity trade. An additional insight gained from the two-good variant is that the direction of physical capital flow may hinge on the identity of the factor that is used intensively in the industry with random technology.

◆

References

Anderson, James E. "The Heckscher-Ohlin and Travis-Vanek Theorems under Uncertainty." *J. Internat. Econ.* 11 (May 1981): 239–47.

Batra, Raveendra N. "Production Uncertainty and the Heckscher-Ohlin Theorem." *Rev. Econ. Studies* 42 (April 1975): 259–68.

Eaton, Jonathan, and Gersovitz, Mark. "A Theory of Expropriation and Deviations from Perfect Capital Mobility." Working Paper no. 972. Cambridge, Mass.: N.B.E.R., August 1982.

Helpman, Elhanan, and Razin, Assaf. *A Theory of International Trade under Uncertainty.* New York: Academic Press, 1978. (*a*)

———. "Uncertainty and International Trade in the Presence of Stock Markets." *Rev. Econ Studies* 45 (June 1978): 239–50. (*b*)

Kemp, Murray C., and Liviatan, Nissan. "Production and Trade Patterns under Uncertainty." *Econ. Record* 49 (June 1973): 215–27.

MacDougall, G. D. A. "The Benefits and Costs of Private Investment from Abroad. A Theoretical Approach." *Econ Record* 36 (March 1960): 13–35.

Mundell, Robert A. "International Trade and Factor Mobility." *A.E.R.* 47 (June 1957): 321–35.

Rothschild, Michael, and Stiglitz, Joseph E. "Increasing Risk. I. A Definition." *J. Econ. Theory* 2 (September 1970): 225–43.

Svensson, Lars E. O. "Factor Trade and Goods Trade." *J. Internat. Econ.* (forthcoming).

Turnovsky, Stephen J. "Technological and Price Uncertainty in a Ricardian Model of International Trade." *Rev. Econ. Studies* 41 (April 1974): 201–17.

Are Economic Time Series Asymmetric over the Business Cycle?

Salih N. Neftçi

Graduate School, City University of New York

It has long been argued that major cyclical variables such as the unemployment rate display an asymmetric behavior over various phases of the business cycle. The paper provides a statistical test for this hypothesis. Using the framework of finite state Markov processes I implement a test to see if the behavior of the unemployment rate is characterized by sudden jumps and slower drops. It is argued that the framework provided in the paper can also be used to test other sample path properties of economic time series.

This paper investigates the issue whether correlation properties of major economic time series differ across various phases of the "business cycle"—a property denoted by "asymmetry." The pro treated using the statistical theory of finite-state Markov p and the applications are limited to various unemploymen these being a reasonable measure of the business cycle.

The claim that major economic time series are asymm different phases of the cycle arises in almost all major works ness cycles. For example, according to Keynes (1936, p. 314), substitution of a downward for an upward tendency often takes place suddenly and violently, whereas there is, as a rule, no such sharp turning point when a upward is substituted for a downward tendency" (see also Hicks 1950). In figure 1 I illustrate a well-known

I wish to thank Richard Porter, Gary Shinasi, and Christopher Sims for valuable discussions of an earlier version of this paper. A referee's comments are greatly appreciated.

[*Journal of Political Economy*, 1984, vol. 92, no. 2]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0006$01.50

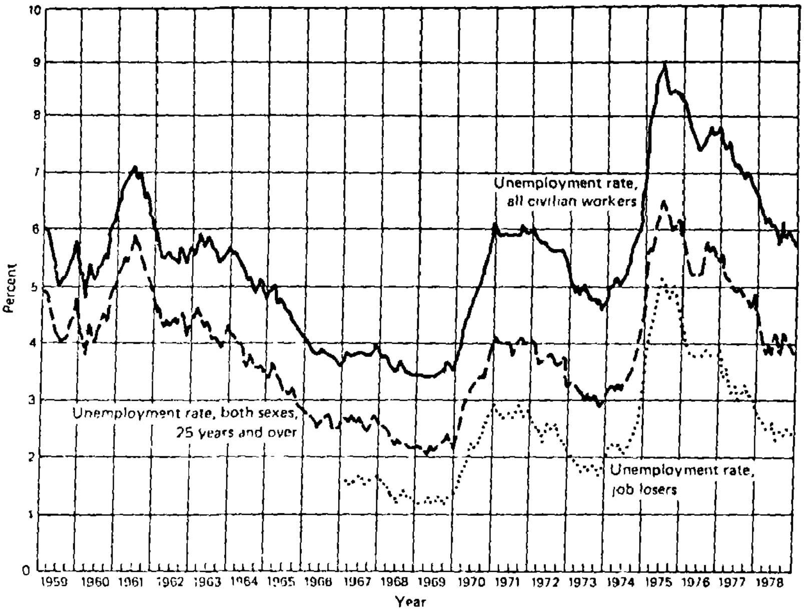


FIG. 1

example where increases in the unemployment rate are much sharper than declines.

The question, then, is whether these asymmetric drops occur systematically enough to be counted as part of the probability structure of economic time series. It is important to detect the presence of such asymmetry for several reasons. First, if asymmetric behavior is indeed systematic then one needs to develop theoretical models that can generate such behavior endogenously. To the best of my knowledge no such model exists at the moment. Further, if economic times series exhibit an asymmetric behavior over the business cycle, then a model that generates sharp drops during contractions followed by gradual movements during expansions will have "better" predictive power. Otherwise, one would expect the "fit" to deteriorate around turning points. For example, as shown in figure 2, the employment equation of the Federal Reserve Board's MPS model seems to imply that the model failed to capture the asymmetric behavior of aggregate employment during the turning points.

The issue of asymmetry is also relevant to the empirical work on rational expectations models. For example, Sargent (1978) and Sargent and Hansen (1979) discuss how dynamic optimization problems of economic agents result in demand and supply relations' depend-

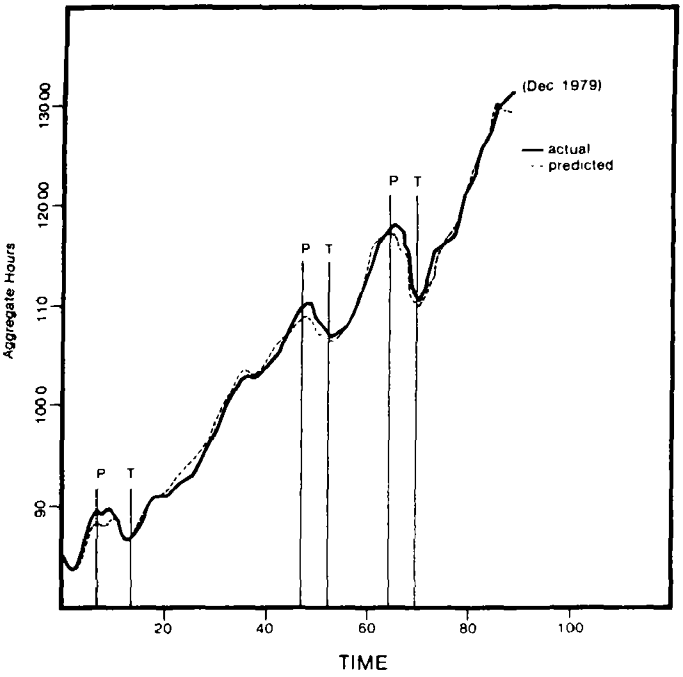


FIG. 2.—Actual and predicted hours, 1958. *QI* – 1

ing, among other things, on the forecasts of future exogenous variables. The existence of an asymmetric behavior in major variables be forecast—for example, unemployment rates and futu rates—could then make these demand and supply relatio tion of some sort of switching regime. Under these condi "prediction problem" will become a nonlinear one, and t nontrivially affect the empirical results (see Neftçi 1982 ample, if a nonlinear prediction problem is treated (by mista linear one, then the estimate of the "unpredictable components" of time series would contain too much information. Under such circumstances, the estimated innovations are more likely to turn up as "significant" variables in an empirical model (Neftçi 1983).

An asymmetric behavior in the business cycle or in interest rates may also account for the belief held by market participants that the prediction of a "turning point" is somehow different from the prediction of these series during "other" times. Finally, the issue of asymmetry is relevant to the way one applies asymptotic econometric theory. As suggested in Hannan (1974), one version of the second-order asymptotic theory depends on the assumption that the best predictor also be the best linear predictor. But, if there is some evidence that nonlinear ARMA models produce predictions with lower mean squared errors due to asymmetry in economic time series, then this would imply some caution in the way asymptotic theory is being used.

The paper is organized as follows. I first discuss how sample path properties of economic time series can be investigated. Then I outline a theoretical framework where the issue of symmetry versus asymmetry is reduced to restrictions on the transition probabilities of the appropriate time series. Next I discuss calculation of confidence ellipsoids and posterior-odds ratios involving these restrictions and report my results. The last section presents my conclusions.

Investigating Sample Path Properties of Economic Time Series

In this section I shall first discuss how the issue of asymmetry should *not* be handled. Let $\{X_t(\omega), t = 0, 1, 2, \ldots; \omega \in \Omega\}$ be a stochastic process defined on the probability space (Ω, F, P), where $\omega \in \Omega$ are events determining particular realizations of $X_t(\omega)$. The notion of asymmetry is usually formulated in terms of the sample path properties of the $X_t(\omega)$ with $\omega \in \Omega$ fixed. For example, this is the case if one compares the behavior of $X_t(\omega)$ when there is a tendency to "decline" with the same series when it shows a tendency to "rise." This comparison is made for a fixed ω, across different t's. According to this, notions such as asymmetry are defined using the sample function of the process under investigation (e.g., see Keynes's or Hicks's comments above). This may bias a test toward the alternative hypothesis. For example, in testing for asymmetry over the business cycle, one may be tempted to divide the sample into two groups, one representing downturns, the other expansions; estimate two separate autoregressive representations for each subsample; and then, using any standard test, see if the two sets of estimates are significantly different from each other. It turns out that, because the definition of asymmetry uses a sample path property of the series under consideration, this test will be biased: by dividing the sample into two groups using a sample property of $\{X_t(\omega)\}$, one increases the probability that the test will reject the null hypothesis—which in this case will be symmetric behavior.

In this paper I argue that the "correct" strategy in investigating hypotheses related to sample path properties of economic time series is to avoid altogether the use of Chow test–type procedures and instead utilize the statistical theory of finite state Markov processes. In particular, I claim that hypotheses relating to sample path properties may be reduced to restrictions on transition probabilities from one state to another and tested using the likelihood function. For example, given a time series $\{X_t\}$, one can define a finite state Markov process $\{I_t\}$ by letting one state represent increases in $\{X_t\}$ and the other represent decreases. Then using the process $\{I_t\}$ one can calculate various transition probabilities and form likelihood ratio tests of symmetry versus asymmetry. This would illustrate to what degree the behavior of $\{X_t\}$ is homogeneous across different states (or regimes).

The methods used in this paper are closely related to panel data models discussed in Heckman (1981*a*) and can easily be implemented to other hypotheses concerning sample path properties of economic time series.

The Theoretical Framework

Let $\{X_t\}$ be a "pro-cyclical," linearly regular stationary process with zero mean, and let $\{I_t\}$ be given by

$$I_t = \begin{cases} +1 \text{ if } \Delta X_t > 0 \\ \\ -1 \text{ if } \Delta X_t \leq 0. \end{cases} \tag{1}$$

The process $\{I_t\}$ will then in general be positive during upswings and negative during downswings. Also $\{I_t, I_{t+s}, s > 0\}$ would normally be correlated.

My claim is that the pattern in which $\{I_t\}$ moves from one state to another can be exploited to investigate any asymmetric behavior in a time series. For example, if the claim is that the observed series show sharp drops during contractions but that during upswings occur gradually, then one would expect $\{I_t\}$ to remain in the longer than the state -1, everything else being the same clearly implies that the transition probability associated wi from $+1$ to $+1$ is greater than the probability associated transition from -1 to -1. Such restrictions implied by th metry can then be used to implement an empirical investigation

In order to obtain a nontrivial theory that we can implement empirically, we have to impose further restrictions on the process $\{I_t\}$. Since my empirical work deals with unemployment rates, and since these

---

[1] Note one advantage of going from $\{X_t\}$ to $\{I_t\}$. The latter variable, being a function of changes in X_t only, will be invariant to monotonic transformations of X_t. Thus, taking logarithms of X_t would yield an identical $\{I_t\}$ process. This suggests that by logging a time series one would not be able to eliminate the kind of asymmetry discussed in this paper.

series exhibit little trend if any, I first assume that $\{I_t\}$ is stationary. Second, I assume that it is a second-order Markov process.[2]

Now let $S_T = \{i_1, i_2, \ldots, i_T\}$ denote a realization of the process $\{I_t\}$. Then we can write the likelihood function as $L(S_T) = P(I_1 = i_1, \ldots, I_T = i_T)$. The assumption that $\{I_t\}$ is second-order Markov makes this

$$L(S_T) = P(I_T = i_T | I_{T-1} = i_{T-1}, I_{T-2} = i_{T-2}) \ldots$$
$$P(I_3 = i_3 | I_2 = i_2, I_1 = i_1) P(I_2 = i_2, I_1 = i_1).$$

After denoting the transition probabilities by $\{\lambda_{ij}, i = 0, 1, \text{and } j = 0, 1\}$ and the probability of initial state by π_o, we can rewrite the likelihood function as

$$\begin{aligned} L(S_T, \lambda_{ij}, \pi_o) = {} & \pi_o(\lambda_{11})^{n_{11}}(1 - \lambda_{11})^{T_{11}}(\lambda_{00})^{n_{00}}(1 - \lambda_{00})^{T_{00}} \\ & \times (\lambda_{01})^{n_{01}}(1 - \lambda_{01})^{T_{01}}(\lambda_{10})^{n_{10}}(1 - \lambda_{10})^{T_{10}}, \end{aligned} \tag{2}$$

where n_{ij}, T_{ij} are the number of observed occurrences of the respective transitions throughout the sample period and where the transition probabilities $\{\lambda_{ij}\}$ are defined as

$$\begin{aligned} \lambda_{11} &= P(I_k = 1 | I_{k-1} = 1, I_{k-2} = 1) \\ \lambda_{00} &= P(I_k = -1 | I_{k-1} = -1, I_{k-2} = -1) \\ \lambda_{10} &= P(I_k = 1 | I_{k-1} = 1, I_{k-2} = -1) \\ \lambda_{01} &= P(I_k = -1 | I_{k-1} = -1, I_{k-2} = 1) \end{aligned} \tag{3}$$

with $\pi_o = P(I_2 = i_2, I_1 = i_1)$. Accordingly, for a model where I_t is second-order Markov, the likelihood function will depend on the four independent parameters $\{\lambda_{ij}\}$ and on the initial condition π_o. Thus, given a sample S_T, the parameters of interest can be estimated in a straightforward manner once we decide how to treat π_o (see also Kedem 1980).

Parameterization of π_o

There are two conditions under which one can treat π_o as a nuisance parameter and to some extent ignore it. These conditions do not apply in this case. However, to motivate parameterization of π_o I will briefly summarize them. One of these cases occurs when the sample size T is "large." As shown by Billingsley (1961), under that condition information contained in π_o on the parameters λ_{ij} will become negligible as $T \to \infty$. In the second case, one can assume that the process I_t

---

[2] This latter assumption can be generalized to kth-order processes in a straightforward manner.

indeed started at time $t = 1$. Under that condition there will be no previous history to I_1, and consequently π_o will not contain any useful information on the λ_{ij}. It turns out that none of these assumptions is reasonable in our case. I am working with basically seven to eight business cycles, so my "observations" are not many. Second, $t = 1$ is clearly not the true starting point of the I_t process. All these imply that π_o may indeed contain useful information on the λ_{ij} that needs to be taken into consideration.

Heckman (1981*b*) provides an example of how one can parameterize π_o. Several options are discussed. In the present case I assume that the time series we are dealing with is stationary (although possibly nonlinear) and has been in operation for quite a while, so that π_o can be replaced by its limiting expression. In fact, in the Appendix I show that for a second-order Markov process the π_o corresponding to the event $\{I_2 = 1, I_1 = 1\}$ will be given by

$$\pi_o = \lim_{N \to -\infty} P(I_2 = 1, I_1 = 1 | I_N = i_N),$$

which reduces to (see the App.)

$$\pi_o = \frac{[(1 - \lambda_{00})(\lambda_{10})]}{[(1 - \lambda_{11})(1 - \lambda_{00} + \lambda_{01})] + [(1 - \lambda_{00})(1 - \lambda_{11} + \lambda_{10})]}. \quad (4)$$

The expressions for π_o relating to initial conditions other than $\{I_2 = 1, I_1 = 1\}$ can be obtained in a similar manner. In any case, expression (4) above suggests that π_o will carry useful information regard parameters $\{\lambda_{ij}\}$. It turns out that when one has relatively fe tions this information may be nonnegligible and may affe mates significantly. Obviously, this is an advantage of usin sion (4) in approximating π_o. However, this is accompli certain cost: calculation of the estimates $\{\hat{\lambda}_{ij}\}$ will now have t through iterative procedures instead of by solving the first-o ditions in a straightforward manner.

The likelihood function will be used in two ways. First I obtain the maximum likelihood estimates (MLEs) of the transition probabilities, λ_{00} and λ_{11}, which describe the way I_t moves during downswings and upswings, respectively. Then, using this likelihood function, I obtain joint confidence intervals for the estimates $\hat{\lambda}_{11}$ and $\hat{\lambda}_{00}$. These will show to what extent the behavior of the time series is different during the two phases of the business cycle. Next, I introduce some reasonable priors and then calculate a posterior-odds ratio for symmetry versus asymmetry.

However, before I discuss these procedures, two issues relative to the data set used in the paper should be dealt with.

Data

I report empirical results for alternative choices of $\{X_t\}$: the unemployment rate (seasonally adjusted), the unemployment rate 15 weeks and over, and the unemployment rate for insured workers. I assume that these series measure roughly what economists call the "business cycle."

There are at least two reasons why the applications were limited to unemployment series. First, I believe that series related to the production side give a better indication of the business cycle. Also, sudden changes in such variables are more conceivable than in such economic series as consumption (see, e.g., Hicks 1950). The consumption series is expected to be "smoother," as the permanent-income hypothesis would suggest. Second, the variables I selected are more or less trend free. This facilitates the implementation of the procedures described below, since I do not need to estimate and eliminate a trend component in order to get the empirical results. However, it is an advantage of the procedures described here that they can handle nonstationary X_t as well, given that the implied I_t will often plausibly be stationary even when X_t is not.[3]

The choice of the observation interval is also nontrivial. The unemployment rate allows me to use either monthly or quarterly data. The discussion in this paper is in terms of quarterly data, for the following reason: Suppose that the time series under consideration has an asymmetric behavior, where gradual upward movements are followed by sudden and sharp drops. Next, suppose we impose on this series a white-noise error having a "small" variance. The resulting series would still "look" asymmetric. But because the upward movements have a small derivative, the runs in $+1$'s will be broken more often than the runs in -1's, which have a bigger derivative. Monthly data would indeed impose such a white noise on these series. However, with quarterly data, white-noise errors are more likely to be averaged out. Thus quarterly data are more likely to show asymmetry. Of course, this is even truer of annual unemployment rate data, which are clearly asymmetric. But, with few observations, the power of the statistical procedures goes down. Using quarterly data is thus a compromise.

Also, a second-order Markov condition may be closer to the truth with quarterly data. In contrast, for monthly time series one may need to introduce further parameters to capture dependencies across states.[4]

---

[3] Notice that since on a priori grounds we expect $\lambda_{00} > \lambda_{11}$, any positive trend in the unemployment rate would bias λ_{11} "toward" λ_{00} and would *enhance* any empirical evidence against symmetry displayed by our methods.

[4] All data were obtained from the Business Conditions Digest (BCD). Using their

Estimates of $\{\lambda_{ij}\}$ and Relevant Confidence Regions

The likelihood function shown in (2) will first be used to obtain estimates for the transition probabilities and then to derive the relevant confidence ellipsoids. To get the estimates, one differentiates the log likelihood function with respect to the $\{\lambda_{ij}\}$:

$$\frac{\partial \log L(S_T, \lambda_{ij}, \pi_o)}{\partial \lambda_{ij}} \propto \frac{\partial \pi_o}{\partial \lambda_{ij}} + \pi_o \frac{n_{ij}}{\lambda_{ij}} - \pi_o \frac{T_{ij}}{(1 - \lambda_{ij})}, \qquad (5)$$

$$i = 0, 1, j = 0, 1,$$

where the first term on the right-hand side is also a function of λ_{ij} but is going to be kept in that compact notation for convenience (the App. contains the exact expression for $\partial\pi_o/\partial\lambda_{ij}$). The estimates $\hat{\lambda}_{ij}$ can be obtained by setting (5) equal to zero. The matrix of second partials denoted by H_T can be obtained in a straightforward manner:

$$\left[\frac{\partial^2 \log L(S_T, \lambda_{ij}, \pi_o)}{\partial \lambda_{ij} \partial \lambda_{kp}}\right] = H_T, \quad i, j, k, p = 0, 1. \qquad (6)$$

In the Appendix we report the exact expression for H_T in terms of the estimated λ_{ij}. It is well known that $-H_T^{-1}$ will be an (approximate) variance-covariance matrix for the MLEs of λ_{ij}. This permits the calculation of the following confidence ellipsoid:

$$[\lambda - \hat{\lambda}]'(-H_T)[\lambda - \hat{\lambda}] = \chi^2_k(\alpha), \qquad (7)$$

where λ is the vector of λ_{ij}'s and $\hat{\lambda}$ is its estimate, and α is a confidence level to be chosen.

Once the transition probabilities are estimated, one can th late the matrix of second partials and plot the confiden shown in (7). This is implemented for the series of unem rates discussed earlier. The center of these ellipsoids will be imum likelihood estimates of the two parameters of inter λ_{11}. All points within the ellipsoids may represent the true λ with a probability of α.

Of special interest is the location of the diagonal with respect to the confidence ellipsoids shown in the figures. The diagonal basically represents the locus of λ_{ij} for which we have $\lambda_{00} = \lambda_{11}$. In other words, all points on the diagonal correspond to equal transition probabilities associated with consecutive declines and consecutive increases. In our terminology the diagonal corresponds to the symmetry. All points off

---

system, unemployment data were computed from series 37, "number of persons unemployed," and series 441, "total civilian labor force." Average weekly insured employment was series 45. Unemployment 15 weeks and over was series 44. All data were for the period 1948–81, except series 45, which was for period 1949–81. Since these data contain frequent revisions and since these revisions *do* make a slight difference, I should also state that the data were obtained from the 1981 issues of the BCD the diagonal relate to an asymmetric behavior. In fact, the more distant the point is from the diagonal the more asymmetric the series will be.

Results discussed in the remainder of the paper use the restriction $\lambda_{00} = \lambda_{11}$ to model the hypothesis of asymmetry. However, with a larger data set and longer time series one can test stronger versions of the symmetry. For example, the restriction $1 - \lambda_{10} = 1 - \lambda_{01}$ may be interpreted as an additional condition of symmetry and incorporated into the null. This was not done in the paper mainly because the literature on business cycles seemed to deal more with the asymmetry of upturns and downturns.

Results

Figures 3–5 display various two-dimensional confidence intervals for the transition probabilities λ_{00} and λ_{11}. The diagonal is the set of $\{\lambda_{ii}, i = 0, 1\}$ where we have $\lambda_{00} = \lambda_{11}$; that is, the series is symmetric. Any other point indicates a situation where $\lambda_{00} \neq \lambda_{11}$, which in turn implies that the times series exhibits a different behavior over periods of persistent declines or upturns. Confidence regions are calculated using the expression shown in (7). These confidence regions are

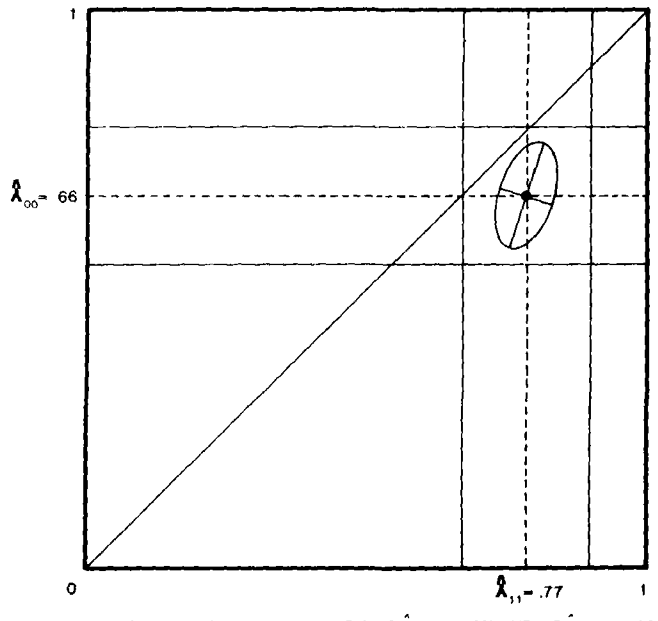


FIG. 3.—Total unemployment rate; SE of $\hat{\lambda}_{11}$ = .07; SE of $\hat{\lambda}_{00}$ = .09

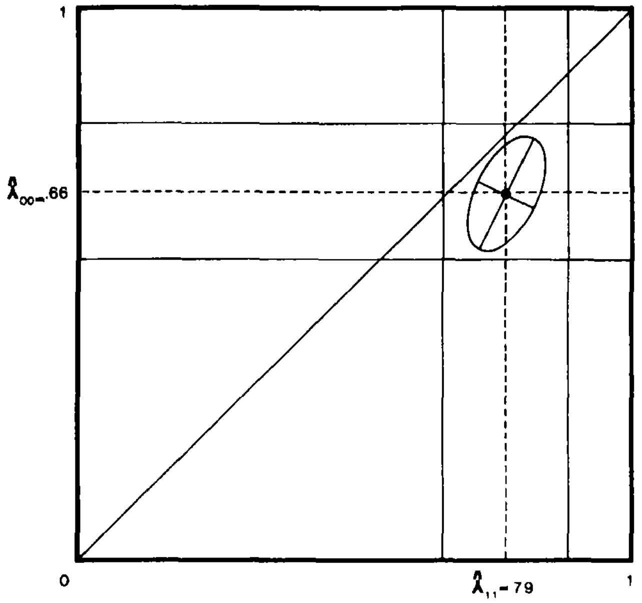


Fig. 4.—Unemployment for insured workers; SE of $\hat{\lambda}_{11}$ = .05, SE of $\hat{\lambda}_{00}$ = .07

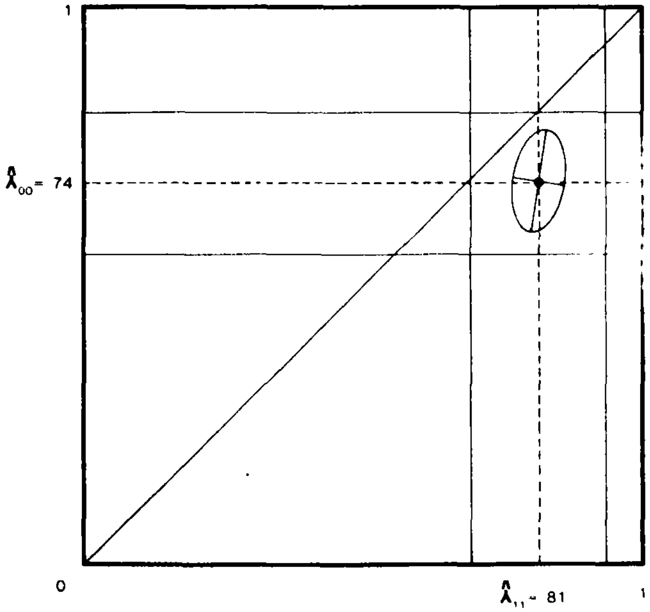


Fig. 5.—Unemployment 15 weeks and over, SE of $\hat{\lambda}_{11}$ = .05; SE of $\hat{\lambda}_{00}$ = .08

(slightly) tilted ellipsoids where the center is the maximum likelihood estimates of λ_{00}, λ_{11}. These estimates are obtained after solving the first-order conditions shown in (5) and (6) iteratively.

Once this is done one can calculate the confidence ellipsoids displayed in figures 3–5 using the matrix of second partial H_T. In each case we see a similar pattern. The diagonal remains outside an 80 percent confidence ellipsoid, providing some evidence against the equality of λ_{00} and λ_{11} and in favor of asymmetric behavior in the observed series. This is more pronounced in the case of insured unemployment rates and in the case of unemployment over 15 weeks. These series are obviously more "cyclical" than the total unemployment rate. However, even in the case of the total unemployment rate, the evidence favors asymmetry.

A Summary of the Evidence: Posterior-Odds Ratios

Confidence regions discussed above are indeed one way of dealing with the evidence concerning any asymmetric behavior in time series. This, however, can be further sharpened, for two reasons. First, it appears that the belief in the asymmetric behavior of unemployment rates is quite widespread, so one can try to impose some reasonable priors on these confidence regions and see how results are affected. Second, confidence regions summarize the evidence only up to a certain point, and the use of such priors may further sharpen the results by permitting calculation of a posterior-odds ratio. This statistic can then be interpreted as summarizing the evidence involving symmetry versus asymmetry. In this section I describe how this posterior-odds ratio can be calculated.

First note that if $L(S_T, \lambda_{ij})$ represents the likelihood function and $R(\lambda_{ij})$ denotes a reasonable prior, then the posterior probability associated with $\{\lambda_{ij}\}$ and given the sample S_T will be

$$P(\lambda_{ij}|S_T) = \frac{L(\lambda_{ij}, S_T)R(\lambda_{ij})}{P(S_T)}, \tag{8}$$

where $P(S_T)$ is the (unconditional) probability of obtaining the sample S_T and is an unknown constant; $P(\lambda_{ij}|S_T)$, on the other hand, can be interpreted in a standard fashion as the probability associated with various values of the transition probabilities λ_{ij}, conditional on the sample and on the priors used.

Clearly, if one is interested in comparing two opposing views concerning the λ_{ij} one can calculate respective probabilities and form an odds ratio.

To be more precise, I let the four-dimensional unit rectangle be denoted by R^4. We then note that $\{\lambda_{ij}\} \in R^4$. Suppose H_a and H_o are two disjoint subsets of R^4 representing two alternative hypotheses or points of view. Then the posterior-odds ratio can be calculated as follows:

$$r(H_a, H_o, S_T) = \frac{P(\lambda_{ij} \in H_a | S_T)}{P(\lambda_{ij} \in H_o | S_T)}, \tag{9}$$

where $P(\cdot | S_T)$ is the a posteriori probability defined by (8).

In our case the set H_a will represent those values of λ_{ij} that lead to an asymmetric behavior of the time series under consideration. H_o, on the other hand, will represent symmetry. Here we let H_o be those values for λ_{ij} for which $\lambda_{00} = \lambda_{11}$. In figures 3–5 this set was shown as the diagonal. Thus we let

$$H_o = \{\lambda_{00}, \lambda_{11}: \lambda_{00} = \lambda_{11},\ 1 > \lambda_{01} > 0,\ 1 > \lambda_{10} > 0\}, \tag{10}$$

and H_a is the complement of H_o in R^4. The posterior-odds ratio will then be given by

$$r(H_a, H_o, S_T) = \frac{\int_{\lambda_{ij} \in H_a} p(\lambda_{ij} | S_T) dp}{\int_{\lambda_{ij} \in H_o} p(\lambda_{ij} | S_T) dp}, \tag{11}$$

where $p(\cdot | S_T)$ is the posterior density.

According to the expression in (11), to calculate the posterior odds we need to (1) select a reasonable conjugate prior on λ_{ij}, (2) obtain the posterior density $p(\cdot | S_T)$, and (3) calculate the probabilities associated with H_a and H_o. Some of these are nontrivial steps, so before reporting the empirical results I will discuss these items.

Selection of the Priors

To see how we can proceed in selecting a prior, consider th hood function again:

$$\begin{aligned} L(S_T, \lambda_{ij}) = {} & C(\lambda_{01}, \lambda_{10}) \Phi(\lambda_{00}, \lambda_{11}, \lambda_{01}, \lambda_{10}) \\ & \times \lambda_{00}^{n_{00}} (1 - \lambda_{00})^{r_{00}} \lambda_{11}^{n_{11}} (1 - \lambda_{11})^{r_{11}}. \end{aligned} \tag{12}$$

Obviously, this function depends on the critical parameters λ_{00} and λ_{11} through two different terms. The first term $\Phi(\lambda_{00}, \lambda_{11}, \lambda_{10}, \lambda_{01})$ is there because of the initial condition probabilities previously denoted by π_o. The exact expression for $\Phi(\cdot)$ was derived in the Appendix and was shown to be a complicated function of the λ_{ij}. As can be seen from the expression in the Appendix, this term does not have a nice functional form. The second term involving the λ_{00} and λ_{11}, on the other hand, looks like a multivariate beta distribution. This immediately suggests the "beta family" as the natural conjugate priors. Thus, we now assume that the prior judgments that one has on λ_{00} and λ_{11} can be summarized as a bivariate density:

$$f(\lambda_{00}, \lambda_{11}) = B(a_0, b_0)B(a_1, b_1)\lambda_{00}^{a_0}(1 - \lambda_{00})^{b_0}\lambda_{11}^{a_1}(1 - \lambda_{11})^{b_1}, \quad (13)$$

where the constants $B(a_i, b_i)$ are given by:

$$B(a_i - 1, b_i - 1) = \frac{(a_i + b_i - 1)!}{(a_i - 1)!(b_i - 1)!}, \quad i = 0, 1. \quad (14)$$

Finally, the likelihood function depends on the λ_{01} and λ_{10}, which we treat as constants and denote for convenience by $C(\lambda_{01}, \lambda_{01})$.

In finalizing the selection of the prior we should keep in mind the purpose of our investigation. Basically, we would like to compare the "mass" of posterior probability that can be assigned to the "events" $H_a = \{\lambda_{00} \neq \lambda_{11}\}$ and $H_o = \{\lambda_{00} = \lambda_{11}\}$. Now under the conditions outlined thus far, the set that consists of $\lambda_{00} = \lambda_{11}$ will have a measure zero for any strictly continuous prior. For this reason, in finalizing the selection of the prior, we have to use some discrete probabilities associated with H_a and H_o. Accordingly, we combine the bivariate beta density with discrete probabilities to get the prior $R(\lambda_{00}, \lambda_{11})$, as follows:

$$R(\lambda_{00}, \lambda_{11}) =$$

$$\begin{cases} (1 - \alpha_0)\lambda_{00}^{n_{00}}(1 - \lambda_{00})\lambda_{11}^{n_{11}}(1 - \lambda_{11})^{T_{11}} & \text{if } \lambda_{11} \neq \lambda_{00} \\ \alpha_1\lambda^{n_{00} + n_{11}}(1 - \lambda)^{T_{00} + T_{11}} & \text{if } \lambda_{11} = \lambda_{00} \end{cases} \quad (15)$$

with $(1 - \alpha_0) = (1 - \alpha)B(a_0, b_0)B(a_1, b_1)$ and $\alpha_1 = \alpha B(a_0 + a_1, b_0 + b_1)$, and where the $0 < \alpha < 1$ is a nonzero "spike" representing the prior probability of symmetry.

Calculation of the Posterior-Odds Ratio

Given the prior density in (15) and the likelihood function in (2) one can calculate the posterior-odds ratio involving symmetry versus asymmetry by integrating the posterior over appropriate sets:

$$r(H_a, H_o, S_T) = \frac{\int_{\lambda_{00} \neq \lambda_{11}} p(\lambda_{00}, \lambda_{11}|S_T)dp}{\int_{\lambda_{00} = \lambda_{11}} p(\lambda_{00}, \lambda_{11}|S_T)dp}. \quad (16)$$

However, given the presence of $\Phi(\lambda_{00}, \lambda_{11}, \lambda_{10}, \lambda_{01})$ in the likelihood function, the integration indicated in (16) will be computationally forbidding. But a good lower bound for $r(\cdot)$ can be easily calculated by a simple transformation.

It turns out that the presence of $\Phi(\cdot)$ in (2) is due to the initial conditions in the sample S_T. In the empirical work we are dealing with, these initial conditions are always given by $\{I_2 = +1, I_1 = +1\}$. Also, I_0, although unobserved, can be either $+1$ or -1. But this means that the initial sequence can be either $\{+1, +1, +1\}$ or $\{+1, +1, -1\}$. Clearly this is a restriction on all possible sequences, and that is why $\Phi(\cdot)$ is informative on λ_{11} and λ_{00}. In the empirical work presented below we shall assume that the initial sequence is the one where $\{I_2 = +1, I_1 = +1, I_0 = -1\}$. Obviously this will bias the posterior-odds ratio $r(\cdot)$, but only slightly. At the same time it leads to the elimination of the term $\Phi(\cdot)$, while yielding a reasonable estimate for the posterior-odds ratio. If the use of quarterly data eliminates most of the transient white noise from the unemployment series, then this estimate will constitute a good lower bound for $r(\cdot)$. If, on the other hand, the sharpness of upturns in unemployment is more likely to overwhelm transient white noise than is the subtlety of long downturns, then our estimate will be an upper bound. Indeed, this is one reason why quarterly data were used. It is clear from figure 1 that with *monthly* data the prior expectation $\lambda_{11} > \lambda_{00}$ is more appropriate. Under such conditions, however, a lower bound will be obtained by choosing $\{I_2 = +1, I_1 = +1, I_0 = -1\}$ as the initial sequence. With quarterly data, we hope that most of the transient noise that exists in monthly series will be eliminated. Thus we pick $\{I_2 = +1, I_1 = +1, I_0 = +1\}$ as the initial sequence and obtain the approximate posterior-odds ratio:

$$
\begin{aligned}
r(H_a, H_o, S_I) = {} & \left[(1 - \alpha_0)\int_0^1\int_0^1 (\lambda_{00})^{n_{00} + a_0}(1 - \lambda_{00})^{T_{00} + b_0}\right. \\
& \left. \times (\lambda_{11})^{n_{11} + a_1 + 1}(1 - \lambda_{11})^{T_{11} + b_1} d\lambda_{11} d\lambda_{00}\right] \\
& \div \left[\alpha_1 \int_0^1 (\lambda)^{n_{00} + a_0 + n_{11} + a_1 + 1}(1 - \lambda)^{T_{00} + b_0 + T_{11}}\right.
\end{aligned}
$$

This can be simplified by using the fact that beta densities integrate to one:

$$
\left[\int_0^1\int_0^1 (\lambda_{00})^{n_{00} + a_0}(1 - \lambda_{00})^{T_{00} + b_0} (\lambda_{11})^{n_{11} + a_1 + 1}(1 - \lambda_{11})^{T_{11} + b_1} d\lambda_{00} d\lambda_{11}\right]
$$
$$
\times B(n_{00} + a_0, T_{00} + b_0)B(n_{11} + a_1 + 1, T_{11} + b_1) = 1
$$

and

$$
\left[\int_0^1 (\lambda)^{n_{00} + a_0 + n_{11} + a_1 + 1}(1 - \lambda)^{T_{00} + b_0 + T_{11} + b_1} d\lambda\right]
$$
$$
\times B(n_{00} + a_0 + n_{11} + a_1 + 1, T_{00} + b_0 + T_{11} + b_1) = 1.
$$

We then obtain a concise expression for the posterior-odds ratio:

$$r(H_a, H_s, S_T) = \frac{1 - \alpha_0}{\alpha_1} \cdot \frac{B(n_{00} + a_0 + n_{11} + a_1 + 1, T_{00} + b_0 + T_{11} + b_1)}{B(n_{00} + a_0, T_{00} + b_0)B(n_{11} + a_1 + 1, T_{11} + b_1)}.$$

Using this expression and the sufficient statistics $\{n_{ii}, T_{ii}, i = 1, 0\}$ we can calculate several estimates for the posterior-odds ratio using various values for the parameters $\{a_i, b_i\}$ of the prior density shown in (15).

Results concerning posterior-odds ratios are shown in table 1, which displays odds ratios for symmetry versus asymmetry using three different priors. In the first case, shown in the first column, I assume that the parameters a_i, b_i, $i = 0, 1$, in the beta priors are all zero. The prior then reduces to two spikes. All λ_{ii}, $i = 0, 1$, so that $\lambda_{00} \neq \lambda_{11}$ will carry a prior probability $(1 - \alpha)$. The values of λ_{ii}, $i = 0, 1$, so that $\lambda_{00} = \lambda_{11}$ will have a flat probability α (see eq. [13] describing the general formula for beta priors). I selected $\alpha = ⅓$ for the first column of table 1. This selection gives a prior-odds ratio of 2 and is certainly fairly conservative. The posterior-odds ratios are then calculated as 0.93 for the total unemployment rate and 1.40 for insured unemployment.

The next two columns use priors that are more favorable to asymmetry. To be more exact, these should be called "empirical priors" since the selected a_i, b_i, $i = 0, 1$, were not solely the result of subjective feelings about λ_{00} and λ_{11} but instead were determined using yearly prewar data on unemployment rates. Column 3 in table 1 selects a_i, b_i so as to give a prior mean for λ_{00}, λ_{11} that coincides with the prewar estimates of these parameters. Column 2 preserves the same point estimates but assumes that they were obtained using half as many observations. Posterior-odds ratios calculated with these priors clearly favor asymmetry much more in the case of the total unemployment rate and the unemployment rate for insured workers. In fact, column 3 of table 1 is like calculating odds ratios using data from the prewar period. For these columns the spike in the priors, namely the parameter α, was still selected as ⅓. This gives a prior-odds ratio equal to 2 and consequently suggests that very little prior information was imposed on the results obtained in the first three columns.

The last two columns are the same sample odds multiplied with a prior-odds ratio equal to 3. In essence, these two columns are the ones that impose any real prior beliefs in asymmetry. The parameter α is selected as ¼, which suggests that on a priori grounds we regard asymmetry as the more likely outcome. Combining these prior odds with the sample we get the posterior odds. As can be seen from the

TABLE 1

POSTERIOR-ODDS RATIOS

		EMPIRICAL BETA PRIORS WITH $\alpha = 1/3$		POSTERIOR ODDS WITH A PRIOR-ODDS RATIO EQUAL TO 3 $(\alpha = 1/4)$	
UNEMPLOYMENT VARIABLE	FLAT PRIORS WITH $\alpha = 1/4$ $a_i, b_i = 0$ $i = 0, 1$	$a_0 = 10$ $a_1 = 18$ $b_0 = 5$ $b_1 = 5$	$a_0 = 20$ $a_1 = 36$ $b_0 = 10$ $b_1 = 10$	$a_0 = 10$ $a_1 = 18$ $b_0 = 5$ $b_1 = 5$	$a_0 = 20$ $a_1 = 36$ $b_0 = 10$ $b_1 = 10$
Total unemployment	93	2.25	2.76	5.03	6.20
Insured	1.40	3.25	3.92	7.30	8.78
Fifteen weeks and over	64	1.48	1.79	3.34	4.02

table the strongest case in favor of asymmetry turns out to be for insured unemployment. Table 1 suggests that asymmetry may be seven times more likely than symmetry. The total unemployment rate also has a high posterior-odds ratio equal to 6.2. The posterior-odds ratios are the lowest for unemployment of 15 weeks and over. These stand at 4.02.

The results offer some evidence in favor of the asymmetric behavior of the time series considered in this paper. In the next section I discuss the implications of such asymmetry.

Implications of Asymmetry

The evidence concerning asymmetric behavior of macroeconomic variables has implications for building statistical or—preferably—economic models that may characterize observed asymmetry. In this section I briefly discuss a series of questions raised by the empirical work reported in this paper.

The first implication of the asymmetry from the point of view of economic theory is that models should be built that can explain this common phenomenon endogenously.

Second, the kind of asymmetry seen in the "cyclical" variables studied here suggests that the probabilistic structure of these series may indeed be different during upswings and downswings. Optimization problems of economic agents need to be reformulated so that they explicitly take into account the existence of such switches. As an example, a producer who observes a set of exogenous variables and who maximizes the expected value of some (quadratic) objective function under standard assumptions would desire to obtain linear forecasts of the exogenous variables. The resulting decision rules (demand or supply schedules) would then be functions of these forecasts. However, if exogenous variables exhibit some asymmetry, then the decision maker may also desire to forecast the periods during which observed processes change their stochastic behavior in addition to estimating conditional means. Under these conditions linear forecasts will fail to be "optimal." This will have important implications for the empirical work involving rational (optimal) expectations. Neftçi (1983) deals with an example that shows how nonlinear forecasts may significantly differ from linear ones when observed series contain random, discrete jumps.

The major econometric question related to the results reported above has to do with how one would represent asymmetric series using time-series models. For example, consider a linear moving average representation of $\{X_t\}$ given by Robinson (1978):

$$X_t = \epsilon_t + a_1\epsilon_{t-1} + a_2\epsilon_{t-2}\epsilon_{t-1}. \tag{17}$$

This model is a nonlinear moving average representation and will, in general, generate different forecasts for upswings and downswings. For example, note that the realizations $\{\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}\}$ and $\{-\epsilon_t, -\epsilon_{t-1}, -\epsilon_{t-2}\}$ would yield, respectively,

$$\hat{X}^1_{t+1} = a_1\epsilon_t + a_2\epsilon_t\epsilon_{t-1}$$

and

$$\hat{X}^2_{t+1} = -a_1\epsilon_t + a_2\epsilon_t\epsilon_{t-1},$$

where $\hat{X}^1_{t+1}$ and $\hat{X}^2_{t+1}$ are forecasts that relate to upswings and downswings, respectively. If a_1 is positive and a_2 is negative, then the model in (17) would yield sharp drops and gradual increases in $\{\hat{X}_{t+1}\}$.

Thus, one way to represent asymmetric series is to use nonlinear time-series models. But these models are cumbersome to estimate and do not have easily tractable properties such as invertibility. Thus, it may be of some practical importance to see if a *linear* model, where the innovations have asymmetric densities, can achieve the same objective. It turns out that the kind of symmetric behavior over the business cycle that I have discussed *can* be represented by a linear model where the filter $\{a_i\}$ may be linear but the ϵ_t's may have an asymmetric density. For computational as well as theoretical reasons, it is important to be able to distinguish between these two possibilities Obviously, further work on this point is needed.

A related issue is the use of asymptotic theory. To the ext economic processes can be more closely approximated by n moving average representation as in (17), the best linear pr will be different from the best predictor (in both the mean an sense). But as Hannan (1974) argues, the "best linear predict the best predictor" and is exactly what gives the asymptotic tion theory commonly used in dynamic models. This suggests th kind of asymmetry discussed in this paper may demand some care in the way one uses standard F- and t-tests in the presence of asymmetry

Conclusions

In this paper, I investigated asymmetric behavior of economic time series. The procedure compares the sample evidence of consecutive declines and consecutive increases in a linearly regular, ergodic pro-

cess and is based on the idea that in a "symmetric series" with no deterministic components, the behavior during upswings and downswings would be similar, in the sense that transitions within one state (decline) would equal transitions within the other state (jumps).

The procedure is applied to different employment-related series. Overall, the results suggest that there is enough evidence in the data to warrant a serious consideration of whether economic time series indeed go through two different regimes during the business cycle.

Appendix

I. Calculation of the Asymptotic Expansion for π_{ij}

In the case of a second-order Markov process the initial probability π_{ij} will relate to four possible events: $\{I_2 = +1, I_1 = +1\}$, $\{I_2 = -1, I_1 = -1\}$, $\{I_2 = -1, I_1 = +1\}$, and $\{I_2 = +1, I_1 = -1\}$. From these I will show how to get limiting probabilities for $\pi_{11} = P(I_2 = +1, I_1 = +1)$ and $\pi_{00} = P(I_2 = -1, I_1 = -1)$. Other limiting probabilities can then be obtained from these in a straightforward manner.

Using the fact that the sequences $\{+1, +1, +1\}$ and $\{+1, +1, -1\}$ represent disjoint events, we can write:

$$\begin{aligned} P(I_2 = +1, I_1 = +1) = {} & P(I_2 = +1 | I_1 = +1, I_0 = +1) \\ & \times P(I_1 = +1, I_0 = +1) \\ & + P(I_2 = +1 | I_1 = +1, I_0 = -1) \\ & \times P(I_1 = +1, I_0 = -1) \end{aligned}$$

and similarly for $P(I_2 = -1, I_1 = -1)$:

$$\begin{aligned} P(I_2 = -1, I_1 = -1) = {} & P(I_2 = -1 | I_1 = -1, I_0 = -1) \\ & \times P(I_1 = -1, I_0 = -1) \\ & + P(I_2 = -1 | I_1 = -1, I_0 = +1) \\ & \times P(I_1 = -1, I_0 = +1). \end{aligned}$$

In matrix form we have:

$$\begin{bmatrix} \pi_{11}(t) \\ \pi_{00}(t) \\ \pi_{10}(t) \\ \pi_{01}(t) \end{bmatrix} = \begin{bmatrix} \lambda_{11} & 0 & \lambda_{10} & 0 \\ 0 & \lambda_{00} & 0 & \lambda_{01} \\ 0 & (1 - \lambda_{00}) & 0 & (1 - \lambda_{01}) \\ (1 - \lambda_{11}) & 0 & (1 - \lambda_{10}) & 0 \end{bmatrix} \begin{bmatrix} \pi_{11}(t - 1) \\ \pi_{00}(t - 1) \\ \pi_{10}(t - 1) \\ \pi_{01}(t - 1) \end{bmatrix}$$

We rewrite this as $\Pi(t) = A\Pi(t = 1)$. By repeated substitution we get $\Pi(t) = A^N\Pi(t - N)$. As $N \to \infty$:

$$\lim \Pi(t) = \Pi^*.$$

This gives us the system of steady-state equations

$$\Pi^* = A\Pi^*, \tag{A1}$$

which we can solve for π_{ij} by using the condition $\Sigma\pi_{ij} = 1$. The resulting asymptotic probabilities turn out to be

$$\pi_{11} = \frac{(1 - \lambda_{00})\lambda_{10}}{(1 - \lambda_{11})(1 - \lambda_{00} + \lambda_{01}) + (1 - \lambda_{00})(1 - \lambda_{11} + \lambda_{10})}$$

$$\pi_{00} = \frac{(1 - \lambda_{11})\lambda_{01}}{(1 - \lambda_{00})(1 - \lambda_{11} + \lambda_{10}) + (1 - \lambda_{11})(1 - \lambda_{00} + \lambda_{01})},$$

which can be used to calculate π_{10} and π_{01}:

$$\pi_{01} = \frac{(\lambda_{10} - \lambda_{00})}{(2 - \lambda_{10})}\pi_{00} + \frac{(1 - \lambda_{10})}{(2 - \lambda_{10})}(1 - \pi_{11}),$$

$$\pi_{10} = 1 - \pi_{11} - \pi_{00} - \pi_{01}.$$

II. First and Second Partials of the Likelihood Function

The likelihood function $L(S_T, \lambda_{ij})$ was given by

$$L = \pi_o \lambda_{00}^{n_{00}}(1 - \lambda_{00})^{T_{00}} \lambda_{11}^{n_{11}}(1 - \lambda_{11})^{T_{11}} \lambda_{01}^{n_{01}}(1 - \lambda_{01})^{T_{01}} \lambda_{10}^{n_{10}}(1 - \lambda_{10})^{T_{10}}.$$

Below we report first and second partials of the logarithm of this function:

$$\frac{\partial \log L}{\partial \lambda_{00}} \propto \frac{\partial \pi_o}{\partial \lambda_{00}} + \pi_o \frac{n_{00}}{\lambda_{00}} - \pi_o \frac{T_{00}}{(1 - \lambda_{00})} = 0,$$

$$\frac{\partial \log L}{\partial \lambda_{11}} \propto \frac{\partial \pi_o}{\partial \lambda_{11}} + \pi_o \frac{n_{11}}{\lambda_{11}} - \pi_o \frac{T_{11}}{(1 - \lambda_{11})} = 0,$$

$$\frac{\partial \log L}{\partial \lambda_{01}} \propto \frac{\partial \pi_o}{\partial \lambda_{10}} + \pi_o \frac{n_{10}}{\lambda_{10}} - \pi_o \frac{T_{10}}{(1 - \lambda_{10})} = 0,$$

$$\frac{\partial \log L}{\partial \lambda_{10}} \propto \frac{\partial \pi_o}{\partial \lambda_{01}} + \pi_o \frac{n_{01}}{\lambda_{01}} - \pi_o \frac{T_{01}}{(1 - \lambda_{10})} = 0,$$

where, for the initial state $\{I_2 = +1, I_1 = +1\}$.

$$\pi_o = \frac{(1 - \lambda_{11})\lambda_{01}}{2(1 - \lambda_{00})(1 - \lambda_{11}) + (1 - \lambda_{00})\lambda_{10} + (1 - \lambda_{11})\lambda_{01}}$$

$$= \frac{(1 - \lambda_{11})\lambda_{01}}{(1 - \lambda_{00})(1 - \lambda_{11} + \lambda_{10}) + (1 - \lambda_{11})(1 - \lambda_{00} + \lambda_{01})},$$

$$\frac{\partial \pi_o}{\partial \lambda_{00}} = \frac{[(1 - \lambda_{11} + \lambda_{10}) + (1 - \lambda_{11})]\pi_o}{(1 - \lambda_{00})(1 - \lambda_{11} + \lambda_{10}) + (1 - \lambda_{11})(1 - \lambda_{00} + \lambda_{01})},$$

$$\frac{\partial \pi_o}{\partial \lambda_{11}} = \frac{-1}{(1 - \lambda_{11})}\pi_o + \frac{\pi_o[(1 - \lambda_{10} + \lambda_{01}) + (1 - \lambda_{00})]}{(1 - \lambda_{00})(1 - \lambda_{11} + \lambda_{10}) + (1 - \lambda_{11})(1 - \lambda_{00} + \lambda_{01})},$$

$$\frac{\partial \pi_o}{\partial \lambda_{01}} = \frac{\pi_o}{\lambda_{01}} - \frac{(1 - \lambda_{11})\pi_o}{(1 - \lambda_{00})(1 - \lambda_{11} + \lambda_{10}) + (1 - \lambda_{11})\lambda_{01}},$$

$$\frac{\partial \pi_o}{\partial \lambda_{10}} = \frac{-\pi_o(1 - \lambda_{00})}{(1 - \lambda_{00})(1 - \lambda_{11} + \lambda_{10}) + (1 - \lambda_{11})(1 - \lambda_{00} + \lambda_{01})}.$$

The second partials of the likelihood function will be given by

$$\frac{\partial^2 \log L}{\partial \lambda_{00}^2} = -\frac{n_{00}}{(\lambda_{00})^2} - \frac{T_{00}}{(1-\lambda_{00})^2} + \frac{[(1-\lambda_{11}+\lambda_{10})+(1-\lambda_{11})][(1-\lambda_{11}+\lambda_{10})+(1-\lambda_{11})]}{[(1-\lambda_{00})(1-\lambda_{11}+\lambda_{10})+(1-\lambda_{11})(1-\lambda_{00}+\lambda_{01})]^2},$$

$$\frac{\partial^2 \log L}{\partial \lambda_{11}^2} = -\frac{n_{11}}{(\lambda_{11})^2} - \frac{T_{11}}{(1-\lambda_{11})^2} - \frac{1}{(1-\lambda_{11})^2} + \frac{[(1-\lambda_{00}+\lambda_{01})+(1-\lambda_{00})]^2}{[(1-\lambda_{00})(1-\lambda_{11}+\lambda_{10})+(1-\lambda_{11})(1-\lambda_{00}+\lambda_{01})]^2},$$

$$\frac{\partial^2 \log L}{\partial \lambda_{00}\partial \lambda_{11}} = \frac{[(1-\lambda_{11}+\lambda_{10})+(1-\lambda_{11})][(1-\lambda_{00}+\lambda_{01})+(1-\lambda_{00})]}{[(1-\lambda_{00})(1-\lambda_{11}+\lambda_{10})+(1-\lambda_{11})(1-\lambda_{00}+\lambda_{01})]^2} - \frac{2}{[(1-\lambda_{00})(1-\lambda_{11}+\lambda_{10})+(1-\lambda_{11})(1-\lambda_{00}+\lambda_{01})]}.$$

These partials are then used in forming the matrix H_T discussed in the text.

References

Billingsley, Patrick. *Statistical Inference for Markov Processes.* Chicago: Univ. Chicago Press, 1961.

Hannan, E. J. "Time Series Analysis." *IEEE Trans. Automatic Control* 19 (December 1974): 706–15.

Heckman, James L. "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time–Discrete Data Stochastic Process." In *Structural Analysis of Discrete Data with Econometric Application,* edited by Charles F. Manski and Daniel McFadden. Cambridge, Mass.: MIT Press, 1981. (*a*)

———. "Statistical Models for Discrete Panel Data." In *Structural Analysis of Discrete Data with Econometric Application,* edited by Charles F. Manski and Daniel McFadden. Cambridge, Mass.: MIT Press, 1981. (*b*)

Hicks, John R. *A Contribution to the Theory of the Trade Cycle.* Oxford: Clarendon Press, 1950.

Kedem, Benjamin. "Estimation of the Parameters in Stationary Autoregressive Processes after Hard Limiting." *J. American Statis. Assoc.* 75 (March 1980): 146–53.

Keynes, John Maynard. *The General Theory of Employment, Interest and Money* London: Macmillan, 1936.

Neftçi, Salih N. "Optimal Prediction of Cyclical Downturns." *J. Econ. Dynamics and Control* 4 (August 1982): 225–41.

———. "Expectations of Variables Containing Random Discrete Jumps." Manuscript. New York: City Univ. New York, 1983.

Robinson, P. M. "Estimation of a Nonlinear Moving Average Process." *J. Stochastic Processes* 3 (1978): 135–56.

Sargent, Thomas J. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86 (December 1978): 1009–44.

Sargent, Thomas J., and Hansen, Lars P. "Formulating and Estimating Dynamic Linear Rational Expectations Models." Manuscript. Minneapolis: Univ. Minnesota, 1979.

Confirmations and Contradictions

The Incidence of a Tax on Pure Rent: The Old Reason for the Old Answer

George Fane

Australian National University

The purpose of this note is to clarify the formal identities linking the problems analyzed by Barro (1974), Feldstein (1977), and Calvo, Kotlikoff, and Rodríguez (1979) (henceforth CKR) and to reaffirm the traditional proposition that in both the short run and the long run a tax on pure rent is *never* shifted, provided one adopts the traditional procedure in tax theory of analyzing the incidence of a compensated tax. Any uncompensated tax is equivalent to a compensated tax together with a lump-sum redistribution of income, and there nothing "surprising" about the fact that lump-sum redistribu income may alter relative prices, as the literature on the problem" in international trade theory demonstrates. Actual tions can be found (e.g., Gorman 1953; Barro 1974; Calvo et p. 874, n. 7) in which lump-sum redistributions of income affect relative prices. Even when they do, the incidence effect small relative to the size of the redistribution.

The analyses of CKR and Feldstein give the impression classical proposition on the incidence of a tax on pure rent d apply in a dynamic economy: "As Feldstein points out, the classical proposition that a tax on pure rent is unshifted requires that the supplies of nonland factors of production are unaltered by the introduction of the tax on land rents. This static assumption of fixed factor supplies is inappropriate to the analysis of tax incidence in a dynamic economy in which the supply of capital reflects the economic choice between consumption and saving" (Calvo et al. 1979, p. 869).

[*Journal of Political Economy*, 1984, vol. 92, no. 2]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9202-0011$01.50

In fact, the proposition that a compensated tax on pure rent is not shifted does apply in a dynamic economy and does not require any static assumption of fixed supplies of nonland factors. In all the cases analyzed by Feldstein and CKR in which the tax on pure rent is shifted, the shifting occurs only because these authors never analyze the case of a fully compensated tax on pure rent. An uncompensated tax on pure rent will of course generally be shifted in a static model, as in a dynamic model.

Barro (1974) analyzed an overlapping-generations model in which each generation's utility depends on the utility of their children to a sufficient extent to induce positive bequests from parents to children. In this model a switch from financing government expenditure by taxing the current generation to financing it by bond sales to the current generation, with the principal and interest to be financed by increased taxation of future generations, has no effect on the levels of consumption or utility of any generation. The switch represents a transfer by the government to parents from their children, which the parents can and would exactly offset by raising the bequests they make to their children.

Feldstein (1977) analyzed the effects of a tax on land rents in an overlapping-generations model in which bequests are not made. The young work, consume, and save; their savings finance the acquisition of land and capital, the earnings from which are used to finance consumption in retirement. Feldstein assumed that the revenue from the land rent tax is wasted (or, equivalently, "used to finance some additional government expenditure that does not increase utility vis-à-vis the no-tax situation" [Feldstein 1977, p. 349]) and showed that in general some of the tax will be shifted. That is, the introduction of a land rent tax at rate θ will not in general reduce the value of land to a fraction $(1 - \theta)$ of its original value. In fact, the value of land may even rise.

CKR changed two of Feldstein's assumptions: first, they assumed that the revenue from the tax on the rent of the land held by the older generation is not wasted but returned period by period either to the younger generation or to the older. Second, they analyzed this amended tax in the context of a Barro-type overlapping-generations model *with* bequests and no inheritance taxes. They showed that under these revised assumptions the tax will *not* be shifted.

CKR also noted that Feldstein had been wrong to assert (Feldstein 1977, p. 352, n. 10) that the tax will not be shifted in his no-bequest model if the revenue is returned to the older generation: "While we [i.e., CKR] discuss the case of a compensated tax on land rents in which the government returns the tax revenues in a lump-sum fashion to either the young or old, Feldstein analyzes an uncompensated tax on land rents. *However, his result for his model holds for a compensated tax as well.* The traditional practice in incidence analysis is to consider compensated taxes" (Calvo et al. 1979, p. 870; emphasis added).

Their comment is misleading; what CKR presumably mean is that if the revenue from the land rent tax, which is collected in each period from the currently older generation (who happen to hold all the land under the particular institutional assumptions of both Feldstein and CKR), is returned in a lump-sum fashion in each period to the currently older generation, the tax will, in general, be shifted to some extent. This is correct, and Feldstein's apparent denial of it was an error, as CKR note. However, it is wrong to describe this case as a compensated tax. To follow "the traditional practice in incidence analysis" of considering compensated taxes, it is necessary to assume that at the time of announcing and introducing the new tax on land rent the government simultaneously issues infinitely dated bonds, uses the land rent tax revenue to meet the interest payments on these bonds period by period, and uses the proceeds from the sale of the bonds to finance a once-over lump-sum handout to the owners of the land on the day that the land rent tax is announced. This procedure fully compensates the original landlords, while the procedure assumed by CKR (of paying the entire land rent tax revenue, period by period, to the older generation) clearly fails to do so.

In the case of a fully compensated tax on land rents, as described above, the entire incidence of the tax will be exactly reflected in a fall in land values, even in Feldstein's no-bequest model. This conclusion, that a compensated tax on pure rent is never shifted, remains true when one considers Feldstein's second reason (namely, the portfolio effect) for asserting that the uncompensated tax that he analyzed will be shifted: the uncompensated tax on land rents changes the
tions by value of land to capital in portfolios, and attempts
holders to restore the initial proportions will generally tend
the price of land. However, this portfolio effect will also be
tive in the case of the fully compensated land rent tax, since th
issued bonds would substitute for the reduced value of land in po
folios.

If land rents are stochastic, full compensation would require the interest payments on the bonds to be stochastic also, since full compensation would require the land rent tax revenue and the interest payments to be exactly equal period by period. Under these arrangements the bonds would be a *perfect* substitute for the reduced value of land in portfolios. An arrangement in which the newly issued bonds offered a sure interest payment each period, even though land rents (and therefore land rent tax revenues) were stochastic, could not be described as full compensation, since the government would need to raise additional revenue (e.g., by a poll tax) in periods when rents were low and would have a surplus to distribute in periods when rents were high. To the extent that the resulting poll taxes and subsidies involved uncompensated intergenerational transfers, the proposition that the land rent tax is not shifted would no longer in general be exactly true. However, in a model with fully operative intergenerational bequests, these uncompensated, stochastic, intergenerational transfers, induced by the government, would be fully offset by changes in intergenerational bequests for reasons that are formally equivalent to those analyzed by Chan (1983).

The cases considered by CKR (and by Feldstein 1977, p. 352, n. 10) are therefore equivalent to (i) the introduction of a fully compensated land rent tax together with (ii) a switch from bond finance to tax finance of the type analyzed by Barro (i.e., the imposition of a lump-sum tax on the particular generation who are old at the time of the introduction of the tax, with the tax revenue being used to finance bond purchases by the government). In every model, both in the short run and in the long-run steady state, the incidence of i falls entirely on land values with no shifting. In CKR's model, with bequests, the transfer involved in ii is exactly offset for the reasons analyzed by Barro. This is the underlying explanation for CKR's "no shifting" result. The no shifting result holds both in the long-run steady state (which CKR analyze) and in the short run (cf. Calvo et al. 1979, p. 874, n. 7), since Barro's analysis is not restricted to the steady state. However, in Feldstein's model, *without* bequests, the transfer involved in ii will generally affect relative prices, thus modifying the incidence of the tax. Even with bequests eliminated, this modification may well be relatively small.

Finally, the analysis above can be viewed as providing an extension of the list of cases in which the Ricardo-Barro proposition (that bonds are not net wealth) is valid. Barro established this proposition for the overlapping-generations model *with* bequests, but it clearly also applies in the model *without* bequests if the principal and interest payments on newly issued bonds are to be met by new taxes on the future rents of assets (e.g., land) currently owned by the individuals who are to benefit from the once-over lump-sum tax cut that the bond sale finances.

References

Barro, Robert J. "Are Government Bonds Net Wealth?" *J.P.E.* 82 (November/December 1974): 1095–1117.

Calvo, Guillermo A.; Kotlikoff, Laurence J.; and Rodríguez, Carlos Alfredo. "The Incidence of a Tax on Pure Rent: A New (?) Reason for an Old Answer." *J.P.E.* 87 (August 1979): 869–74.

Chan, L. K. C. "Uncertainty and the Neutrality of Government Financing Policy." *J. Monetary Econ.* 11 (May 1983): 351–72.

Feldstein, Martin. "The Surprising Incidence of a Tax on Pure Rent: A New Answer to an Old Question." *J.P.E.* 85 (April 1977): 349–60.

Gorman, William M. "Community Preference Fields." *Econometrica* 21 (January 1953): 63–80.

Book Reviews

Overseas Investment in the Age of High Imperialism: The United Kingdom, 1850–1914. By MICHAEL EDELSTEIN.
New York: Columbia University Press, 1982. Pp. 367. $35.00 (cloth); $15.00 (paper).

Britain's export of capital before the First World War has attracted the attention of economists interested in the nineteenth-century growth of trade and international specialization, the gold standard, European imperialism, and the performance of the British economy. The magnitude of the flow of British wealth abroad was unique; from 1870 to 1913 more than 5 percent of British income flowed abroad and about a third of all British assets at the outbreak of the war were foreign claims. Michael Edelstein has made a major contribution to the understanding of the phenomenon. In contrast to the most recent contributions he has focused principally on the British capital market and the British economy and on long-run trends, although he also pays due attention to borrowers' demand for funds and to the long-swing timing of the flows. His analysis leads him to reject views that emphasize biases in the British capital market or a worsening of the British income distribution as major sources of capital exports, while still assigning greater weight to British savings behavior and investment opportunities than would many who have emphasized the growth of the Atlantic economy in their studies of capital flows.

The principal analysis of the study involves an attempt to ascertain the relative importance of British supply and foreign demand factors in the determination of British capital flows. A series of chapters collected under the general heading "Micro Processes" begins with an illuminating examination of the structure of the British capital market with particular reference to the allegation that the export of capital resulted from a bias in the London capital market—the MacMillan Gap. Edelstein rejects this view and argues persuasively that the needs of smaller firms were adequately met at lower cost in regional capital markets. London's concentration on large issues was specialization related to comparative cost and not bias.

This is followed by an econometric investigation of British accumulation of American securities within the context of total British asset accumulation. The conclusion—which is central to the overall conclusion of the study, but not without its problems—suggests that British accumulation played a more important role in determining the annual time series of American issues (principally by railroads) in London than did the timing of American railroad investment. The interesting suggestion is made that by the 1870s American railroad securities possessed return-risk characteristics that made them attrac-

[*Journal of Political Economy*, 1984, vol. 92, no. 2]
© 1984 by The University of Chicago. All rights reserved

tive assets in British portfolios, particularly since new British railroad securities were no longer available. The demand for these American securities arose in proportion to growing British portfolios.

The "Micro Processes" section of the book concludes with an expansion of Edelstein's already published investigation of the realized yields on various first- and second-class domestic and foreign investments issued in London. The ranking of equity and fixed income securities of various types of enterprises shows little surprise over the long run with, generally, debenture yields below equity, public utility and railroad yields below industrials, and home yields below foreign. It is perhaps surprising to find U.S. railway debt with the highest return of any debenture and the return not apparently justified by its variance. This may, however, be an artifact of the selection of first- and second-class securities that eliminates any security that encountered difficulties over the period. It seems unlikely that railroads that were eliminated from the series (e.g., the Reading, Erie, Wabash, Sante Fe, Central Pacific, and Northern Pacific) looked worse ex ante than railroads that did not in fact experience difficulties. The most interesting finding of this investigation of returns is the alternation of the superiority of foreign and home returns that leads the alternation of home and foreign investment. The alternation of realized returns was principally caused by a stronger long cycle in domestic returns than in foreign and is interpreted as evidence to support the view that falling British opportunities were an important cause of capital exports.

The second half of the book is devoted to largely noneconometric investigations of the macro determinants of investment and saving in Britain and the principal borrowing countries. The objective is to proportion the causes in long-term trends of capital exports between shifts in British excess supply of funds and overseas excess demand. Investment demand in Britain and overseas is modeled as consisting of population-sensitive and productivity-sensitive proportions. Savings in turn is seen as a product of the dependency ratio and age structure. The results may be summarized briefly. In Britain savings increased and combined with slower demographic growth and failing industrial productivity, which resulted in only slow shifts in the investment demand schedule, to produce an excess supply of funds. The United
reasonably viewed as the most important of the borrowers. Here in
and 1860s investment demand shifted dramatically relative to do
ings and played a major role in the initial rise of British foreig
Thereafter, however, the supply of domestic savings and investmen
are seen—largely on the strength of the earlier econometrics—to h
together, and so no major pull of capital came from the United S
expansion of smaller economies of recent settlement, particularly Aus
the 1880s and Canada after the turn of the century, is seen as provid
weaker pull forces. Edelstein feels, however, that there is some evidence of excess in these expansions, particularly that of Canada, which he attributes to an extremely liberal source of funding in Britain.

Edelstein's overall conclusion can then be roughly summarized as follows. The initial spurt of lending seems to have been a response to the need for capital in areas of recent settlement, particularly the United States. These pressures did not entirely disappear, but by the late 1870s, British capital seeking employment became the driving force.

There is much in the book that will be welcomed by all those working in the field, but this general conclusion rests on weak enough grounds that those who believe that the expansion of the world economy played a larger role in the capital flow will remain skeptical. The argument revolves principally around the United States, and here the principal evidence is the econometric analysis in chapter 4 in the "Micro Processes" section. Unfortunately this is the most vulnerable portion of the book. The problem is first and foremost one of data choice but also involves the specification of American supply of securities. The excess supply of American securities is modeled as a function of the nominal stock of U.S. railway tangible capital, the total U.S. capital stock (as a proxy for U.S. wealth), and interest rates. The problem here arises from the use of Melville Ulmer's estimate of U.S. railway tangible capital. This series is certainly inadequate (see Fishlow 1966, pp. 589–94) and is particularly suspect in the 1880s. One of the greatest surges of British lending to the United States occurred between 1887 and 1890. These years also represent the most rapid expansion of the U.S. railroad network. The miles of railroads in the United States increased by a quarter in these years, but Ulmer's capital series misses this investment and increases by only 11 percent. The use of the Ulmer series thus seems grounds enough to doubt Edelstein's results. There are in addition at least two additional concerns about this econometrics. First, Edelstein reasonably suggests a portfolio view of British holdings of U.S. securities, but this must surely call into question the adequacy of the new calls series as an indicator of the annual flow. After all *The Times* quoted New York prices on American railroad securities continuously. The second problem here emerges from the failure to consider the U.S. government debt. It is well known that a great deal of U.S. Civil War debt was held in Britain and was rapidly retired; it seems appropriate to try to incorporate this into the modeling.

Edelstein's view of the massive flow of funds to Canada in the early twentieth century as evidence of excess capital pushed from Britain might also be questioned. Edelstein expresses the strong suspicion "that the supply of U.K. savings to Canada was overabundant and in its absence the rate of Canadian railway investment spending would have been considerably smaller" (p. 281). This conclusion is based on casual comparison experience elsewhere and the obvious excess of the third Canadian transcontinental railroad. The Canadian experience, however, does not seem obviously excessive in a North American context. Between 1896 and 1914 the Canadian railroad network was extended some 14,400 miles. Perhaps 2,000 of these miles were the uneconomic duplication by the National Transcontinental in mountains and the Canadian shield; 8,450 were new track in the wheat-producing regions of the Prairie Provinces. A yardstick of comparison might be provided by the previous great expansion that occurred in the American Prairies. Between 1878 and 1893 the rail network had been extended about 36,700 miles in the North Western states of North and South Dakota, Minnesota, Iowa, Missouri, Kansas, and Nebraska, an area of only some two or three times the agricultural areas opened in Canada.

Professor Edelstein has thus produced a very valuable book on the operation of the British capital market and its overseas involvement. His argument that British forces became dominant by the 1880s will command attention. Nonetheless one who believes that the pull of expanding settlement in new regions in response to world population growth and transportation improvement was the key to British capital movement need not fully abandon that view.

C. Knick Harley

University of Western Ontario

Reference

Fishlow, Albert. "Productivity and Technological Change in the Railroad Sector, 1840–1910." In *Output, Employment and Productivity in the United States after 1800,* edited by Dorothy S. Brady. New York: Columbia Univ Press (for N.B.E.R.), 1966.

Money and Inflation. By FRANK HAHN.
Cambridge, Mass.: M.I.T. Press, 1983. Pp. xii + 116. $12.50.

This short book publishes the Mitsui Lectures in Economics that Frank Hahn gave at Birmingham University in 1981. Hahn is mostly concerned with two central problems in monetary theory—the role of money and the effects of monetary policy. He also discusses monetary policy prescriptively. The lectures teach us something about the economy and about economics. On the latter score, they illustrate the ability of a skilled theorist to unravel some of the economy's mysteries, but they also illustrate the limitations of an economist who employs no special expertise beyond formal reasoning

The first theoretical problem involves extending Walrasian equilibrium analysis to accommodate fiat money that has a positive real value. Hahn's discussion of this problem is instructive and includes useful observations about the effects of the existence of money on the real equilibrium of the economy. He stresses that understanding the role of money requires identifying a service that only holding money provides. Accordingly, he argues that the natural strategy is to go beyond Walras by modeling the technology of exchange. Most models of monetary exchange assume, often implicitly, some form of a cash-in-advance requirement. Hahn points out that closing such models requires analysis of the determination of the timing of receipts and payments.

In the course of this discussion, Hahn reviews an example of a divergent rational inflationary bubble, in which the real money stock, assumed to be initially less than its possible positive steady-state value, declines at an increasing rate asymptotically to zero. The context of this example is a rational expectations model of an infinitely lived economy with overlapping generations and with fiat money as the only asset. Hahn's analysis, although clearly presented, neglects two technical considerations that are worth noting [illegible] without an unanticipated change in the structure of the economy, a possibility that seems inconsistent with rational expectations, a rational bubble cannot exist in this deterministic setting unless the real money stock has been less than its steady-state value literally since the creation of the economy—that is, unless such a bubble has always existed. Second, this model, because it does not allow for asset substitutability in portfolios, has the apparently peculiar characteristic that, depending on the elasticity of substitution between present and future consumption, a rational inflationary bubble could be convergent rather than divergent.

This example of an inflationary bubble is worth studying primarily because it helps to clarify some conceptual and logical issues associated with the idea of rational expectations. This part of Hahn's discussion also illustrates his tendency to attach unwarranted practical significance to formal theoretical exercises. "We have exhibited a world with rational expectations, inflation and a *constant stock of money.* Next time you read, probably in a letter to *The Times,* that 'a necessary and sufficient condition for inflation is an increasing stock of money', I hope you will remember this simple result" (p. 13). This suggestion that an abstract possibility of an inflation that does not involve monetary expansion can help to explain contemporary inflation and that it rebuts "monetarist" policy recommendations is incredible. Had Hahn been less anxious and less clumsy about trying to score points, he would have noted that cyclical variation in the actual inflation rate implies that his bubble example does not fit the facts.

The second theoretical problem involves extending or amending Walrasian equilibrium analysis to account for the apparent real effects of monetary policy. Hahn's discussion begins by connecting this problem to the problem of understanding the role of money. He argues convincingly that in any extended Walrasian model that includes a meaningful role for money—that is, in which fiat money has positive real value—changes in the nominal money stock, whether anticipated or unanticipated, affect real outcomes. He stresses that this failure of neutrality obtains in general because, except in special cases, the process of changing the nominal money stock alters the distribution of endowments. An example, frequently discussed in the literature, is the distribution effect that results when the government exchanges fiat money for bonds, if private agents do not fully consolidate public and private assets and liabilities. Anticipated changes in the nominal money stock involve an additional type of nonneutrality associated with changes in the relative anticipated returns from holding money and real assets.

Hahn is ambivalent about the practical importance of the distribution effects associated with monetary policy. He asserts that ignoring these effects is not harmless, but he also acknowledges that his analysis does not lead even to unambiguous qualitative results, especially about the relation between distribution effects and real aggregates.

Turning to considerations that are more clearly relevant to the relation of money to macroeconomic fluctuations, Hahn is highly critical of attempts to understand the apparent effect of monetary policy on the cyclical behavior of real variables within equilibrium models that extend Walrasian analysis only by introducing imperfect information. Hahn points out correctly that such models seem unable to fit the facts that they purport to explain. Most of the usual list of alleged false implications, which Hahn indiscriminately recites, only reflect simplifying assumptions used in some versions of these models. More important, however, Hahn stresses the testable and easily rejected implication of equilibrium models that monetary policy has real effects only if it is unpredictable and unobservable. It is worth noting that Hahn does not mention the possibility, which some equilibrium theorists seem to take seriously, that the observed relation between monetary and real aggregates reflects causation running either from real aggregates or from factors responsible for fluctuations in real aggregates to monetary aggregates. In the quest for a more satisfactory model of the real effects of monetary policy, Hahn, like many others, looks again to Keynes for direction. In the end, however, Hahn does not get effectively beyond a retelling of Keynesian stories.

Hahn mainly emphasizes that failure of markets to clear and resulting Keynesian outcomes are not inconsistent with rational expectations. This correct point, however, does not confront the essential problem involved in understanding how a Keynesian relation between monetary policy and real aggregates can occur. An equilibrium in which the price level is approximately proportionate to the money stock and Keynesian effects are absent has especially appealing properties. The theoretical derivation of such a non-

Keynesian equilibrium follows naturally from neoclassical maximization postulates and the presumption that exchange exhausts gains from trade, ideas that have proven powerful for understanding a wide range of economic phenomena. Moreover, convincing and apparently true stories exist about how prices and quantities respond to supply and demand and drive actual economies eventually to such an equilibrium. The problem for theory, then, is to explain precisely why the behavior of presumably rational agents apparently fails to speed up this convergence sufficiently to remove discernible Keynesian traces from the data.

Hahn's idea, admittedly provisional, is to revive Keynes's assumption that workers are concerned about relative wages. This assumption, however, whether realistic or not, does not seem to direct us toward a satisfactory solution. Most important, an equilibrium with rigid nominal wages would seem to require unanimous resistance to relative wage changes. The behavior of a single deviant worker who did not care about relative wages would start the economy moving toward a non-Keynesian equilibrium. The analysis then must return to the question of speed of convergence. Hahn's inconclusive discussion, as he appropriately notes, reflects how far we still are from a satisfactory theory of macroeconomic fluctuations.

Leaving theory for the moment, the weakest aspects of these lectures are Hahn's attempts to evaluate and prescribe monetary policy. As suggested above, Hahn does not seem to appreciate the inability of formal theorizing alone to generate operational policy prescriptions, and he overestimates the influence of theoretical developments, especially equilibrium macroeconomic models, on policy.

Hahn advocates a straightforward monetary policy. He supposes that there exists excess supply in the labor market and, hence, what he defines as "involuntary unemployment." An expansionary monetary policy can reduce this unemployment. His policy advice, therefore, is to increase the money stock exactly as necessary to eliminate excess supply and to achieve "full employment," but no more than that.

This discussion exemplifies true innocence. Even supposing that Hahn h
the right model, he never mentions the problem of measuring involu
unemployment. Moreover, he attributes the decline in enthusiasm for
monetary policy in recent years to the concurrent shift in theoretical em
away from Keynesian models. He does not consider the possibility that
developments result from the realization by economists and persons
fairs, including Mr. Callaghan and Mrs. Thatcher, that his simplis
proach to monetary policy, right out of a Keynesian textbook, provid
operational bounds on either monetary expansion or inflation.

Another noteworthy shortcoming of these lectures is that Hahn indicates little interest in recent developments in macroeconometrics. Accordingly, his comments about the empirical relevance of theoretical models are limited to judgments about a priori plausibility and to casual and impressionistic observations. More basically, his neglect of macroeconometrics suggests that he does appreciate that the most important contribution that the incorporation of rational expectations into macroeconomic analysis has made to the subject concerns not policy but research strategy. For example, as John Taylor points out in his review of Thomas Sargent's book (*Journal of Monetary Economics*, July 1981), the analysis of rational expectations that follows from the explicit introduction of stochastic elements into the theoretical framework directly generates econometric models that include explicit theoretical specifications of disturbance terms. In this way, the new macroeconomics integrates theory and econometrics.

Returning to Hahn's theoretical discussion, an idea to which he refers with less than characteristic clarity is the theory of implicit contracts. The underlying problem seems to be that, like other theorists who observe the neoclassical dietary laws in their ivory towers, Hahn is not able to suppress completely his discomfort with the *terefah* Keynesian phenomena that the real world seems to serve up. For example, after explaining that the concept of "involuntary unemployment" is meaningful in a Keynesian non-market-clearing context, he suddenly invokes implicit labor contracts, the theory of which, he neglects to mention, is strictly neoclassical and not Keynesian in the least. Then, he writes, "An involuntarily unemployed worker is thus not a 'laid off' worker but one without a contract that he would be willing to accept at terms worse (in utility) than are prevailing" (p. 50). At last, real kosher bacon!

HERSCHEL I. GROSSMAN

Brown University

Low Tax, Simple Tax, Flat Tax. By ROBERT E. HALL and ALVIN RABUSHKA. New York: McGraw-Hill Book Co., 1983. Pp. 137. $18.95 (cloth); $9.95 (paper).

This is an interesting, provocative, and deceptive book. Addressed to the general public and policymakers, it presents a serious proposal for reforming the federal tax system. It is lucidly written and succeeds admirably in avoiding technicalities and tax jargon. In fact, it is so well written that few nontechnical readers will understand the real—and radical—nature of the reform proposal.

The authors propose to substitute for the present federal individual and corporation income taxes a flat rate tax on labor income and unreinvested business income.[1] Labor income is taxed at the individual level; business income that is not reinvested is taxed at the firm level. A personal exemption is allowed only against the tax on labor income. Otherwise, the tax base is comprehensively defined: no deductions are allowed for charitable contributions, interest paid, state and local taxes, or even unusual medical expenses. The result is that the tax return forms are greatly simplified (the authors believe that the individual and business tax forms can be reduced to postcard size, but this is an exaggeration), and problems of tax administration and compliance are reduced. The authors are not only willing, but anxious, to give up rate graduation in the interest of simplicity and improved economic incentives.

The deception arises in the way the authors explain their tax and its effects on the economy. As they tell it, they would tax all income once, just as Irving Fisher proposed almost fifty years ago (Fisher 1937). But what Fisher called an income tax was actually a tax on consumption; so is the Hall-Rabushka tax. Yet the relation between this tax and the consumption tax is mentioned only in two brief paragraphs, two-fifths of the way through the book. To make matters even more puzzling to the reader, there is no explanation of why the tax they propose—essentially a wage tax—is equivalent to a consumption tax.[2] Most people would be appalled by a proposal to substitute a wage tax for the present income tax, so the authors use the income tax label to camouflage the real nature of the tax.

---

[1] The tax is actually a value-added tax of the consumption type. The authors credit E. Cary Brown for providing the theoretical foundations of such a tax in his graduate public finance course, which one of the authors attended.

Comprehensive income taxation has a long history, dating back at least to Henry Simons's *Personal Income Taxation* (1938). The idea is to tax all incomes without any exclusions, personal deductions, or tax credits. The revenue gained in this way would be used to reduce marginal tax rates. It has been estimated that the present average tax burdens in all income classes could be approximated by a comprehensive income tax with a rate schedule ranging from 9 percent in the bottom bracket to 28 percent in the top bracket (Pechman and Scholz 1982).

Hall and Rabushka firmly endorse the comprehensive tax approach. But they would substitute a single rate of 19 percent for the modestly graduated rates in the comprehensive income tax plan. The flat rate of 19 percent is about 40 percent lower than the 28 percent top average marginal rate under present law. The corresponding revenue loss is made up, of course, by higher taxes at lower levels. The break-even point is at an income of about $50,000: taxpayers with incomes above that would pay less tax than they do under present law, while those with lower incomes would pay more.

The authors acknowledge this unfortunate result but argue that the higher income generated by the increased incentives under the flat tax will more than offset the higher tax burden in the lower part of the income scale. They assume that their tax would increase total output in the economy by 9 percent, relying mainly on an estimate by Jerry Hausman of the effect of the present income tax on the number of hours worked. Hausman estimates that replacement of the present income tax by a lump-sum tax would increase total hours of work about 8 percent (Hausman 1982, 1983); the authors assumed that the same estimate would apply to their flat-rate tax. They are aware that the additional 8 percent of hours would yield a smaller add output, so they cut the increase in output to 6 percent. Then, they a assume that there will be a "modest" additional increment to out percent from "drastically improved entrepreneurial incentives." Wh preneurial incentives are improved so much above the present tax which provides many opportunities for tax avoidance, is not expla

The assumption that the flat-rate tax would increase hours of wo percent is incorrect primarily because this estimate applies only to h and omits major segments of the labor force.[3] The flat-rate tax would increase hours of work by more than 8 percent for wives but by less than 8 percent for single parents and unmarried workers. When appropriately weighted, Hausman estimates that the increase in hours of work under a flat-rate tax would be only 5 percent; using the Hall-Rabushka 2 percentage point adjustment for the difference between the responses of hours of work and output, the increase in output would be only 3 percent. Even this may be an overestimate,[4] but to expect an increase twice as large is pure fantasy. Moreover, although the Hausman calculations are measures of the *long-run* effect of lower marginal tax rates, Hall and Rabushka are confident that some effect would be felt immediately and that the full effect will occur by the end of 7 years. Finally, they fail to point out that the more traditional comprehensive income tax plans—those with graduated rates, but at much lower levels than the present schedules—would achieve most of the economic gains of a flat tax.

---

[2] The explanation is as follows. let $Y = C + I = L + B$, where Y is total income, C is consumption, I is investment, L is labor income, and B is business income. The Hall-Rabushka tax base is $L + (B - I) = Y - I = C$. In a world in which individuals consume all their income during their lifetimes and tax rates and interest rates are constant, the present value of L after tax will be equal to the present value of C after tax for each individual (see Warren 1975).

[3] It is also based on a comparison between the present tax system and a lump-sum tax and not a proportional tax. However, this difference is small.

The authors' estimate of the base of their tax is also somewhat cavalier. They derive their estimate from the gross national product data, without allowing for the large differences between incomes reported on tax returns and the national income account figures. Labor incomes are fairly accurately reported on tax returns because of withholding; nonlabor incomes, on the other hand, are underreported by very large amounts (Park 1983). Thus, Hall and Rabushka overestimate their business income tax base by at least $150 billion.[5] To offset this loss of tax base, their flat tax rate of 19 percent would have to be raised to 21 percent.

The advantages and disadvantages of the income and consumption taxes have been debated for years by economists (see Pechman 1980), but Hall and Rabushka have no doubt that the consumption tax is the better approach. They shrug off all the problems of transition, blithely accept the redistribution of tax burdens among individuals and firms entailed by their plan, and—unlike other consumption tax advocates—are completely insensitive to the effects of the consumption tax on the distribution of wealth. Many consumption tax advocates support effective wealth or estate and gift taxes to prevent excessive concentration of wealth and power under a consumption tax. By contrast, Hall and Rabushka would repeal even the weak estate and gift taxes now in effect.

The present income tax is flawed, and it can be greatly improved and simplified. Hall and Rabushka's support for a comprehensive tax is welcome. But their substitution of a consumption tax for the income tax and abandonment of rate graduation would generate a wholesale redistribution of the tax burden that is unwarranted and ill advised.

JOSEPH A. PECHMAN

Brookings Institution

References

Bureau of Economic Analysis. Department of Commerce. "National Income Account Tables." *Surv. Current Bus.* 363 (July 1983): 29.

Fisher, Irving. "Income in Theory and Income Taxation in Practice." *Econometrica* 5 (January 1937): 1–55.

Hausman, Jerry A. "Labor Supply." In *How Taxes Affect Economic Behavior,*

---

[4] See Gary Burtless's comment on Hausman's 1981 paper and James Heckman's comment on his 1983 paper.

[5] Net farm and nonfarm business and partnership profits and rental incomes reported on 1981 individual returns amounted to $51 billion. This compares with a total (excluding imputations) of $203 billion in the national income accounts. Thus, even assuming that the corporate sector reports accurately, Hall and Rabushka overestimate their business tax base by $152 billion. For the tax data, see Internal Revenue Service (1983), and for the national income date, see Bureau of Economic Analysis (1983).

edited by Henry J. Aaron and Joseph A. Pechman. Washington: Brookings Inst., 1981.

———. "Stochastic Problems in the Simulation of Labor Supply." In *Behavioral Simulation Methods in Tax Policy Analysis*, edited by Martin Feldstein. Chicago: Univ Chicago Press (for N.B.E.R.), 1983.

Internal Revenue Service. Treasury Department. *Statistics of Income, 1980, Individual Income Tax Returns*. 1983.

Park, Thai S. "Personal Income and Adjusted Gross Income." *Surv. Current Bus.* 63 (April 1983): 28–33.

Pechman, Joseph A., ed. *What Should Be Taxed: Income or Expenditure?* Washington: Brookings Inst., 1980.

Pechman, Joseph A., and Scholz, John Karl. "Comprehensive Income Taxation and Rate Reduction." *Tax Notes* 17 (October 11, 1982): 83–93.

Simons, Henry. *Personal Income Taxation*. Chicago: Univ. Chicago Press, 1938

Warren, Alvin C., Jr. "Fairness and a Consumption-Type or Cash Flow Personal Income Tax." *Harvard Law Rev.* 88 (March 1975): 931–46

Law and Inflation. By KEITH S. ROSENN.
Philadelphia: University of Pennsylvania Press, 1982. Pp. 451. $10.00

Inflation is usually dealt with as a phenomenon in the realm of economics. Economists are not in agreement as to the causes of inflation, its cures, and even its definition. It is defined, however, in terms used by the economist—money, prices, incomes, and so forth. But the consequences of inflation go beyond the sphere of economic analysis and economic policy, and understanding those consequences requires an interdisciplinary study.

Rosenn's book is a study in the interrelationships between economics and law. It is a rare study of its kind. Despite the fact that all modern economic activity is performed within the legal framework provided by states, economists tend to play down the significance of legal institutions in the sh economic behavior. A similar lack of interest in the working of the ec quite common among lawyers. An eminent legal scholar (Friedma pp. 14–15) explained the disregard often shown by lawyers to econ social realities in the following terms:

> Though the law is obviously concerned with politics, econo cial life, ethics, it is its function to give them form and orde this relation of law to the matters which it regulates, there deve three characteristic features, all sound in principle and indeed fundamental to the idea of law, yet all apt to divorce law from social reality in a time of crisis.
>
> The first of these characteristics is stability . . the desire of the lawyer to preserve stability of conditions may, and often does, blind him to social changes and developments. . .
>
> The second characteristic is formalism . . . there is little doubt that, for the average lawyer, the form of regulating a specific social relation becomes more important than the social relation itself.
>
> Thirdly, and perhaps most important in its effect, there is the desire for security from disorder. . . . Encouraged by the method of legal education, law developed into a skilled craft, a professional technique, pursued with less and less regard for the social matter which it was to regulate.

The consequences of these attitudes are shown in Rosenn's survey of the ways in which various legal systems grappled with the problems of inflation. The nominalist doctrine of the value of money postulates that the value imposed by the sovereign is the value of money. This principle, which was accepted by most civil and common-law systems, embodies the attributes of stability and formalism; it is ill suited to cope with circumstances of persistent inflation But the realities of inflation could not be ignored. In severe cases of runaway inflation, legislative revalorization was sometimes resorted to. Scaling statutes of various kinds were enacted in the United States (following the War of Independence and the Civil War), France (1796–98), Austria (1811), Poland (1924), and Germany (1925). In most cases of inflationary distortions, the task of granting relief from the inequitable consequences was handled by the courts. A whole variety of legal doctrines were devised to enable the modification of contracts undermined by unanticipated inflation. But the prevailing legal doctrine was based on correcting the distortions ex post facto. The introduction of paper money as compulsory legal tender led to the espousal of the nominalist principle as public policy by many legislatures and courts.

The opposition to the use of stabilization clauses designed to protect contracts and other obligations from changes in the value of money is less determined now than it used to be. Rosenn describes the various categories of stabilization clauses and devotes particular attention to the experience of three countries that have instituted indexation on a broad scale—Brazil, Finland, and Israel. No unequivocal policy conclusion emerges from these experiences. Two questions to be answered in the matter of public policy are: (1) Should private parties be free to provide for the protection of their contractual relationships through indexation or similar devices, or should such indexed contracts be considered contrary to public policy? (2) Should the state initiate the introduction of indexation in matters provided for by law: damages, compensation, taxation, and so forth? The case for outlawing indexation of private contracts—which rests on adherence to the nominalist principle—seems very shaky in the light of past experience and present circumstances. But what about adopting comprehensive indexation? Rosenn shows that there is no evidence to support the claim that indexation is likely to become an autonomous source of inflation. It is true that complete wage indexation may lead to inflationary wage settlements, but partial indexation is not subject to the same criticism. It can be clearly shown, on the other hand, that indexation reduces the unfairness and misallocation of resources generated by persistent inflation. The performance of the credit market and of the tax system can be improved by the judicious use of indexation. Rosenn's conclusion is that the benefits outweigh the risks.

Applying my Israeli experience to this conclusion I tend to agree with Rosenn, but with much less confidence than I would have had in its validity only a few years back. Indexation of wage contracts and of financial instruments was introduced in Israel in the late 1940s and the early 1950s. A relatively peaceful atmosphere of industrial relations and a very high rate of personal saving can, at least partly, be attributed to the widespread use of indexation. From 1954 to 1973 the inflation rate in Israel averaged 7 percent per annum—high, in comparison to the United States and Western Europe, but tolerable. In those years the real gross national product was growing at an average annual rate of almost 10 percent.

The pattern changed following the 1973 war and the oil price shock. Economic growth slowed down to less than 3 percent per annum, and the rate of inflation has surged ahead from 30–40 percent a year in 1974–77 to 100–130 percent a year in 1980–82. In an environment of slow growth and rapid inflation, the virtues of overall indexation seem more questionable. On the one hand, the economy has continued to function—severe inflation did not lead to hyperinflation and a collapse of the economic order. But, on the other hand, inflation has fed on itself through the mechanisms of indexation. The political incentives to impose economic measures that may cause unemployment and lower the standard of living were weakened, because wages and savings were protected from erosion by various value clauses. (The impact of inflation has manifested itself in a deteriorating balance-of-payments situation.)

The prolongation of inflation has led to numerous distortions in the allocation of resources. It has necessitated the introduction of complex tax legislation in order to prevent the taxation of capital, on the one hand, and to close various loopholes of tax avoidance, on the other. Indexation has corrected certain evils associated with inflation but has become an impediment to the introduction of economic policies essential to an effective containment of inflation. The Israeli experience of the past decade does not undermine Rosenn's conclusion, but it casts some doubt on its universal applicability.

An interdisciplinary study should be judged according to its usefulness to students of both disciplines. I believe that economists and lawyers alike will benefit from the comprehensive coverage of the subject in this book. An important advantage of *Law and Inflation* is that as a comparative study it provides its readers with a broad range of international experience in an effort to cope with the legal consequences of inflation. Such experience does not offer ready-made solutions to the problems, but it certainly clarifies their complexity. Rosenn's book fills a void in the literature on inflation and should find its place in the libraries of both economists and lawyers.

MEIR HETH

Hebrew University, Jerusalem

Reference

Friedmann, Wolfgang G. *Legal Theory*. 4th ed. London: Stevens, 1960

Economics of Population. By T. PAUL SCHULTZ.
Reading, Mass.: Addison-Wesley Publishing Co., 1981. Pp. 240
(paper).

Although population economics has a venerable history dating at least as far back as Malthus, its latest flowering is relatively recent. The revived interest in human reproduction and household economic activity that began in the mid-1960s has led to a substantial literature on demographic-economic interactions, and Paul Schultz has written an informative review of an important part of it.

Quantitatively, the bulk of recent research on population economics deals with economic and demographic decision making at the household level, and Schultz devotes primary attention to that general area. Indeed, the main organizational theme of the book concerns the contribution that household demand models can make to the study of demographic behavior. Beginning (chap. 2) with a survey of population changes in preindustrial Europe, Schultz uses the Malthusian model as a framework for the analysis of fertility and mortality determinants and as a basis on which to illustrate the potential contributions of the microeconomic theories. Chapter 2 is also notable for its succinct interpretation of recent innovative research by Ronald Lee on the population of preindustrial England.

Chapter 3 is a review of historical and contemporary population trends. It contains sections devoted to trends in population size and distribution, a description of the world mortality decline, the significance of recent age structure shifts in the industrial countries, the impact of rapid population growth in low-income countries, and a discussion of the interaction between technical change and population growth. The latter uses Ester Boserup's theory of population-driven technical change as a vehicle for discussion. Chapter 3 closes with two historical case studies, Taiwan and Colombia. Much of Schultz's own research has been conducted in these two countries, and they serve the same exemplary function throughout the remainder of the book.

The bulk of the remaining chapters is devoted to a presentation of various applications of the new home economics to demographic issues. Chapter 4 is a well-written and useful introduction to the microeconomics of the household that takes the reader from the traditional utility-maximization model, through labor-leisure choices and the time-allocation model of Gary Becker (1965), to multiperson household demand. Particularly valuable is Schultz's discussion in an appendix of the definition and empirical measurement of "full income," a concept from Becker (1965) that has been the source of considerable confusion.

Chapter 5 describes the historical course of mortality trends, reviews the sources of differential mortality, and assesses the role of mortality and morbidity conditions in affecting human reproduction. Schultz is able to draw on his own research on fertility-mortality relationships to inform this chapter.

Chapter 6 surveys economic models of fertility including Easterlin's (1968) explanation of the fluctuations in U.S fertility over the past 5 decades and the distinct line of research inspired by Becker (1960). Easterlin's model is perhaps the most widely cited explanation of the baby boom of the 1950s and of the subsequent "baby bust." Schultz is critical for a number of reasons: (1) Easterlin's theory does not internally generate an explanation of the "secular downtrend in fertility" that characterizes high-income countries. (2) It focuses almost exclusively on supply-side effects on fertility, ignoring changes in aggregate demand for goods and services that might counter the cyclical effects proposed by Easterlin. (3) It pays only token attention to the secular rise in women's labor force participation that has characterized Europe and North America.

Schultz's last point is well founded: Easterlin treats inadequately the steady rise in female labor force participation in the decades following World War II. With respect to the first criticism, however, Schultz misses one of the main features of Easterlin's analysis of U.S. fertility since the mid-nineteenth century. Easterlin takes great pains to argue that the secular decline in U.S. fertility until the mid-1930s can be explained by the decline of the fertility of rural-born native whites and of foreign-born whites. As these two groups became assimilated, their fertility came to resemble that of urban native whites whose fertility had been low since the nineteenth century. It is not, therefore, clear that Easterlin would admit that there is any downward *trend* to explain in postwar American fertility.

This view puts Easterlin in conflict with critics who argue that the current low fertility in the United States is, indeed, the result of a trend that stems irreversibly from social and economic development. The social demographers who hold this view rely almost totally on changes in social norms and tastes in order to predict a permanent decline in aggregate American fertility.

Schultz, however, is reluctant to assign tastes an important role in the determination of demographic behavior. He defines them as residual effects that account for behavioral differences once economic factors are controlled. Then, by arguing that tastes in general are randomly distributed or self-canceling in the aggregate, he dismisses them as determinants of behavior worth study. This is, of course, a common analytical simplification in economics; however, it may not be appropriate in an area in which economic variables generally perform unimpressively in empirical studies.

The second part of chapter 6 is an extended treatment of the microeconomic model of fertility that stems from Becker (1960) and the new home economics time-allocation literature. Schultz presents a lucid account of the technical features of the model and offers useful criticisms. This section is recommended to anyone who wishes to read a succinct and clear assessment of the contributions and problems inherent in the new home economics model of fertility.

Chapter 6 concludes with two empirical case studies of the relationship between fertility and wage rates as well as a discussion of issues in population policy. The latter selection represents the only extended policy discussion in the book, and as such it does not do the topic justice. Population programs, especially as observed in the Third World, offer textbook examples of how economic analysis is applied, misapplied, or not applied to specific policy questions. Schultz could profitably have devoted a chapter to, say, the Indian family planning program, a program that contains a wealth of examples of the use and misuse of economic analysis in population programs.

Chapter 7 is an extensive review of the relationship between lab
participation and economic development. The discussion again reli
household time-allocation model and uses it to analyze secular chang
supply of labor and their relationship to the development process
national comparisons of labor force participation rates are also pr
and Colombia serves as a more detailed case study. This chapter m
benefited from a fuller consideration of the implications of the lab
trends that it documents.

Schultz concludes by considering again how the household-demand model can contribute to the study of demographic behavior. The discussion raises some important issues of research methodology, and not all of Schultz's views are noncontroversial. Because he adopts a model that mandates a complete life-cycle allocation of resources at some well-defined decision point, Schultz is quite emphatic that demands for children, for market work, and for nonmarket time must be characterized as reduced-form demand functions. Therefore, he is very critical of studies that condition, for example, the supply of a mother's market work on the number of children in the household. He argues that both these variables are endogenous and that the reduced-form equation explaining one cannot contain the other as an explanatory variable. Family size and hours of work are, therefore, the jointly determined outcomes of a life-cycle resource-allocation process.

Here Schultz is addressing an issue that is basic to any researcher who uses nonexperimental data: how to represent causal structure and how to divide available variables into endogenous and exogenous sets. The proper way to make these decisions depends fundamentally on the research questions posed and the data available. Use of a particular analytical framework (e.g., the life-cycle resource-allocation model) may be inappropriate given the data available or the specific objectives of the research.

Even if the model is appropriate, use of it does not guarantee that anomalies are avoided. Schultz, for example, is unwilling to accept age at marriage as a predetermined variable in fertility analysis because it too is endogenous in a life-cycle context. Nevertheless, he is willing to award predetermined status to "education" even though it is every bit as likely to be endogenous in a life-cycle model. Issues such as this bedevil nonexperimental data analysis, and empirical researchers cannot as a result afford the inflexible application of theory that Schultz appears to be advocating here.

Perhaps the most disappointing feature of the literature in economic demography as reflected in Schultz's book is its failure to deal seriously with the context within which resource allocation takes place. Economics is primarily the study of allocative decisions made subject to the constraints placed on decision makers by social institutions, governments, resource availability, technology, and prices. Many of these decisions can be studied as if the social system has little systematic effect on them. However, demographic behavior would appear to have strong and systematic connections to the social system, and much of the literature that Schultz has chosen to review suggests that economists have little interest in exploring these connections. This is disappointing because economic demography ought to be more concerned with the response of individual decision makers to changes in the social and economic environment. The reluctance of economists to cast their nets wider diminishes the contribution of the economic approach and hinders its acceptance by noneconomists studying the same questions.

Although *Economics of Population* is, as Schultz is at pains to acknowledge, selective in scope, it is a valuable review of an important part of the revitalized field of economic demography. It provides concise and clear expositions of many important topics and it offers a large but not exhaustive bibliography. Paul Schultz is to be congratulated for contributing an excellent introduction to the economics of population.

Boone A. Turchi

University of North Carolina
Universitat–Gesamthochschule, Paderborn, West Germany

References

Becker, Gary S. "An Economic Analysis of Fertility." In *Demographic and Economic Change in Developed Countries.* Universities–National Bureau Conference Series no. 11. Princeton, N.J.: Princeton Univ. Press (for N.B.E.R.), 1960.

———. "A Theory of the Allocation of Time." *Econ. J.* 75 (September 1965): 493–517.

Easterlin, Richard A. *Population, Labor Force, and Long Swings in Economic Growth: The American Experience.* New York: Columbia Univ. Press (for N.B.E.R.), 1968.

The Informational Efficiency of Experimental Asset Markets

Daniel Friedman

University of California, Los Angeles

Glenn W. Harrison

University of Western Ontario

Jon W. Salmon

First Interstate Bancorp

A fundamental difficulty in devising any test of general efficient market hypotheses is the specification of the relevant public and private information sets of traders. Without a consensus as to reasonable empirical specifications, tests of the hypotheses remain minimal or controversial. In this study we examine the efficiency of experimental asset markets in which one may unambiguously identify the public and private information sets of traders. We focus on how market efficiency is affected by the presence of futures markets in settings that incorporate different types of uncertainty and inside information. Our experimental results support four conclusions: (1) market outcomes tend to evolve toward strong-form informationally efficient equilibria, whether or not futures markets and/or event uncertainty are present; (2) the presence of futures markets clearly stabilizes spot prices; (3) the presence of futures markets tends to speed the evolution of asset markets to more efficient equilibria when there is event uncertainty; and (4) futures markets promote the "leakage" of inside information, with strong-form predictions outperforming semi-strong-form predictions.

We are indebted to the Center for the Study of Futures Markets (Columbia University) and the Foundation for Research in Economics and Education (UCLA) for research support. Len-Huo Hu provided excellent research assistance.

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0011$01.50

I. Introduction

An informationally efficient market may be defined generally as one in which prices and allocations reflect all available information. This definition, coupled with certain assumptions about what might constitute "all available information," leads to testable hypotheses about the behavior of asset markets. A fundamental difficulty in devising any test of such general efficient market hypotheses, however, is the specification of the relevant public and private information sets of traders. Without a consensus as to reasonable empirical specifications, tests of the hypotheses remain minimal or controversial. In this study we examine the efficiency of experimental asset markets in which one may unambiguously identify the public and private information sets of traders. We focus on how market efficiency is affected by the presence of futures markets in settings that incorporate different types of uncertainty and inside information.

A. *Previous Literature*

In recent years theorists have become increasingly interested in the informational role of prices in general and futures prices in particular (see, e.g., Grossman 1976, 1977, 1978; Grossman and Stiglitz 1976, 1980; Bray 1981). These theorists study a world in which information relevant to agents' intertemporal choices is dispersed throughout the economy. Under some conditions they find that competitive equilibrium prices will reveal everything that rational agents need to know to make choices that result in a fully efficient allocation. That is, the economy may achieve an allocation that is the same as that which would arise in a fictitious but otherwise similar economy in which each agent has access not only to his own personal information but also to the personal information of each other agent. In this case prices are said to be perfect aggregators of information and the market is considered strongly efficient. Competitive equilibria reflecting more limited information are referred to below as weakly efficient.[1]

To illustrate the informational role of prices, Grossman (1977) models a case in which spot prices alone do not perfectly aggregate information but spot prices together with futures prices do. He argues that such information aggregation is a major function of futures markets. On the other hand, Svensson (1976), among others, has argued that the existence of sequential trading opportunities and the simultaneous operation of spot and futures markets might lead to Pareto-inefficient outcomes.

---

[1] In Fama (1970), weak-form efficiency has a more precise definition (information sets consist only of past prices) than we employ here. We do employ Fama's distinction between semi-strong-form efficiency (only public information) and strong-form efficiency (all information, public and private, is reflected in prices).

Indeed, there is a widely held belief (long reflected in public opinion, futures markets regulatory proposals, and the writings of economists such as Hardy [1940]) that organized futures markets display many features in common with gambling casinos and induce "unwarranted and undesirable" spot-price fluctuations as volatile futures prices feed back to the spot market via storage and production decisions. Conventional empirical examinations of this controversy involve comparisons of spot-price dispersion for time periods associated with, and then without, the presence of active futures trading in the good. The evidence for this class of tests has been noticeably mixed—see, for example, the results of Cox (1976) and his survey of other studies using this empirical approach. Besides the frequent lack of a theoretical model to tightly specify the tests and interpret the results, the long time periods involved suggest likely violations of the implied ceteris paribus assumption.

The experimental approach offers the possibility of direct tests of market efficiency in which both market structure and individual information sets can be tightly controlled, so its recent popularity in the study of market efficiency issues is not surprising. The studies most relevant to our present concerns are Forsythe, Palfrey, and Plott (1982*a*, 1982*b*), Plott and Sunder (1982), and Friedman, Harrison, and Salmon (1983), hereafter referred to as FPP*a*, FPP*b*, PS, and FHS, respectively. All of these studies employed a double oral auction for multiple units of a single good (the "asset") traded against cas
They differ primarily in their treatment of two critical eleme
asset valuation: time and uncertainty.

PS focused on uncertainty. Their asset expired after a single
but had trader-specific marginal values that depended on an
nous "state of nature." Trade typically occurred before the s
nature was revealed, although in some cases certain agents (
ers") had advance information. They found that equilibrium p
usually revealed the inside information after several repetitions of the market, at least when that information was conclusive as to the true state of nature.

FPP*a* focused on time. Their asset derived its value from cash dividends paid over 2 periods, but each trader knew his own dividend schedule with certainty. Inasmuch as traders did not know others' schedules and therefore did not know at first what prices would be available to them, there existed an endogenous type of uncertainty that we term "market uncertainty" in contrast with the exogenous event uncertainty of PS. FPP*a* found that asset prices approached "rational expectations" values after several repetitions of the market.

The experiments reported in FPP*b*, which were conducted independently of those reported here, extended their previous design in two ways: they systematically examined the role of a futures market and (unlike the rest of the asset market studies, including our own) paid dividends as nonlinear functions of asset holdings. Their main conclusions were that futures markets speeded convergence to rational expectations equilibrium but at the same time increased spot-price volatility during the transition to that equilibrium.

B. Current Approach

FHS was a preliminary report on the first four of a set of nine experiments we conducted at UCLA in 1982.[2] These experiments, discussed more fully in the next section, employed a 3-period design that also allowed for event uncertainty and insider information, with or without futures markets. Within this design we are able to test the robustness of the PS and FPP*a* results, as well as obtain more refined characterizations of informational efficiency than has previously been possible in experimental asset markets.

FHS examined trader experience and the presence of futures markets as treatment variables in a setting with no event uncertainty. FHS confirmed the importance of both variables in speeding convergence toward more informationally efficient prices and found that spot prices were less volatile when futures markets operated. We will comment later on this contrast with the FPP*b* finding.

The present paper is our final report on this set of experiments. In the interest of brevity we will examine below only the six experiments that employed experienced subjects.[3] We omit discussion of learning behavior and other issues raised in FHS that are tangential to informational efficiency. Following a description of our experimental design in Section II, we provide a detailed numerical example in Section III to illustrate the operation of our markets and analytically derive several theoretical equilibria reflecting different amounts of information. We present our experimental results in Section IV and our conclusions in Section V. Appendix A presents more fully the theory underlying our equilibrium concepts and Appendix B provides a detailed descriptive account of individual trader behavior in each experiment.

---

[2] The experiments listed in table 1 below as 1 and 2 appeared in FHS as Experiments 3 and 4 The other experiments listed in table 1 have not been previously reported. We do not report here an experiment using inexperienced subjects that was intended for training purposes

[3] We regard an experiment as involving experienced subjects if at least three-quarters of the current participants have been subjects in a previous 3-period experimental asset market.

II. Experimental Design

A. *Common Experimental Design Features*

The participants in our experiments, referred to as traders, were recruited primarily from MBA classes at the UCLA Graduate School of Management—as likely a habitat for *Homo economicus* as we could think of. After distributing and reviewing the instruction sheets (available on request), we conducted double-oral-auction markets. Our traders were free to announce bid and offer prices and accept the bids or offers of others as long as they did not violate any budget constraint. Transacted prices were publicly recorded. The assets traded were called certificates; they yielded returns, called dividends, to traders who possessed them at the end of each trading period.

Each experiment consisted of a series of "Market Years," which can be thought of as Hicksian weeks. Within each Market Year there were three trading periods, referred to as periods A, B, and C. Each trading period lasted for 5 minutes (real time) and each trader could buy or sell one certificate at a time. Lot sales, short sales (i.e., short spot positions and "naked" short futures positions), and negative cash positions were prohibited.[1] At the beginning of each Market Year each trader was endowed with two certificates and an interest-free loan of $20. The loans were sufficiently large that the liquidity constrai never an impediment to trade.

Incentives for exchange among traders were provided by v the per certificate dividends across individuals as well as acro riods. There were three trader types, with individuals random signed to each group; table 1 provides details of the paramet tions for each experiment. The underlying period-specific certific returns were identical across Market Years—identical in the aggregate *and* for each individual.[2] Thus the markets were repetitively stationary from year to year. Traders were not informed of this stationarity; they had to learn about it in "real time." Each individual was carefully monitored so that his or her private dividend profile was not

---

[1] Plott and Sunder (1983) is the first study to allow short (spot) positions within a trading period. They employ a simple and apparently effective penalty on any trader short at the end of any period (namely, a fixed pecuniary penalty plus the requirement that the trader implicitly cover himself at the highest transaction price during the period).

[2] A minor exception in Market Year 5 of Experiment 2 is noted in table 1

TABLE 1

INDUCED EXPERIMENTAL MARKET PARAMETERS

	AGENTS		DIVIDEND PROFILE					
EXPERIMENT AND MARKET INSTITUTION	Type	ID Number	A		B		C	
1.								
Market uncertainty	I	1, 3, 6	.45		45		45	
No futures	II	4, 7, 9	70		30		.10	
	III	2, 5, 8	10		.30		.70	
2:								
Market uncertainty	I	5, 7, 9	.75		20		10	
Futures	II	1, 3, 6	.40		.45		.45	
	III	2, 4, 8	15		.30		.80	
			X	Y	X	Y	X	Y
3:								
Event uncertainty	I	2, 7, 9	.30	60	25	.53	.35	70
No futures	II	1, 4, 6	15	.15	70	.70	20	20
	III	3, 5, 8	.40	30	53	40	.70	55
4								
Event uncertainty	I	1, 6, 9	25	75	10	60	.50	1 00
Futures	II	3, 5, 8	20	.20	80	80	.30	.30
	III	2, 4, 7	.50	.30	.60	.30	1 00	.70
5:								
Insider uncertainty	I	2, 5*, 6	50	30	.35	30	50	.30
No futures	II	1, 8, 9*	25	25	.45	60	40	20
	III	3, 4, 7*	20	40	30	50	.30	80
6								
Insider uncertainty	I	1, 4*, 9	45	30	35	.35	35	25
Futures	II	3, 5, 7*	.15	50	30	.60	45	30
	III	2*, 6, 8	20	40	.40	.45	.30	70

NOTE – In Market Year 5 of Experiment 2 a random reassignment of agents to trader type occurred, investor type I consisted of agents 1, 3, and 5, type II of agents 4, 6, and 8, and type III of agents 2, 7, and 9. The parameters shown here pertain to all other Market Years. In Experiments 5 and 6 an asterisk beside an agent ID number denotes insiders. *In Market Year 5 of Experiment 5 there was a change in the agents of each type that were insiders agents 2, 1, and 4 were the insiders for that Market Year only.*

observed by any other trader. Possibilities for explicit or implicit collusion were effectively nil.

To motivate the experimental setup one can think of the traders as grain merchants trading in warehouse certificates that have a par value of zero but provide each trader with a finite time profile of convenience yields. Of course, in our experiments traders actually received a cash "dividend" for each certificate held at the end of each trading period, and the certificates expired after the Market Year ended.

In Experiment 1 trading consisted of an immediate exchange of cash for certificates at accepted bid or offer prices (i.e., spot transactions only). In Experiment 2 we permitted futures transactions as well. The futures contract consisted of the delivery of a certificate in period C. Futures contracts as well as spot contracts could be written in both period A and period B. In periods A and B dividends were paid as usual for each (spot) certificate held at the end of that period. No transactions were allowed in period C in Experiment 2, but deliveries previously contracted for were performed. An individual with a net long (short) futures position was required to take (make) delivery of the certificates, and then period C dividends were distributed. A natural interpretation of period C is that it corresponds to the day after the last trading day in the delivery month of a futures contract. Note that an agent had ample opportunity to offset futures positions during periods A and B. However, because of the restriction on short sales, agents' short futures positions were limited to the quantity of inventoried spot certificates at any point in time. For a given net short position an agent's spot sales were also constrained.

In all experiments traders were given a small trading commission of 1 cent per transaction. Such commissions are a standard feature of most experimental market studies; the usual rationale for their inclusion is to overcome subjective transaction costs that might be especially relevant when transacted prices are very close to a market-clearing price. At the end of each experiment we paid our traders in cash for all profits accrued from dividends and trading.

B. The Event and Insider Uncertainty Designs

The distinguishing feature of Experiments 3 and 4 is that eac er's dividend profile is specific to a "state of nature." Two di states, X and Y, were possible. At the beginning of the experin traders were told that we would flip a fair coin at the end o period A trading round, and if the coin came up heads then s occurs and if the coin showed tails then state Y occurs for that M Year. Thus, at the beginning of period A all traders had co prior beliefs about the ultimate realization of state X or Y, based o public knowledge that each state had a 50 percent probability of occurring for each Market Year, and trading during period A occurred under this well-defined event uncertainty. The uncertainty was then resolved as advertised,[6] so periods B and C represented posterior trading rounds in which agents had the opportunity to revise their portfolio of certificate holdings under conditions of *no* event uncer-

---

[6] Plott and Sunder (1983) introduce the idea of actually announcing states of nature from a predetermined sequence. Although there is no evidence that this occurred in their experiments, we were concerned that traders not begin to "second-guess" the sequence of events.

TABLE 2

HYPOTHETICAL EXPERIMENTAL MARKET PARAMETERS

AGENTS		DIVIDEND PROFILE		
Type	ID Number	Period A	Period B	Period C
I	1, 2, 3	.45	.45	.45
II	4, 5, 6	.70	.30	.10
III	7, 8, 9	.10	.30	.70

tainty, as in the previous experimental design. The futures market in Experiment 4 operated in the same manner as in Experiment 2, except that in period A trading for period C certificates was subject to the same event uncertainty as spot trading.

The insider uncertainty design built on the event uncertainty design. The distinguishing feature of Experiments 5 and 6 was that one trader of each trader type was randomly selected to receive conclusive information about which state would occur *before* period A trading took place for each Market Year.[7] This was accomplished by the experimenter's leaving the room before the start of the experiment, flipping a coin, and then writing down the results on 3 out of 9 cards. Cards which read either "no information" or "state _____" were put into envelopes and passed out to the traders. Thus, the six noninsiders did *not* have prior knowledge about the identity of the insiders. These experiments then proceeded in the same manner as the event uncertainty experiments.

III. Theoretical Analysis and Predictions

A. An Example

A concrete numerical example will help the reader develop an intuition for how our markets work and reduce the need for a lengthy abstract discussion in the next few subsections. Let us begin with the hypothetical dividend profile and agent assignment shown in table 2.

Consider the strategic opportunities available to (say) agent 9 in period C—the last trading period of a Market Year with spot trading only. Since he has a dividend profile of type III, he will receive $0.70 for each certificate he owns at the end of the period's trading, and the certificates have no further value to him or anyone else. Hence, $0.70 represents his "reservation price": he will profit by any purchases he can make at a lower price and sales at any higher price. Agent 9 has no direct knowledge of the reservation prices of other agents. He will soon discover, however, that nobody can afford to outbid him, while he cannot undersell anyone. Hence, he may find that if he actively bids and aggressively accepts offers he will acquire certificates at prices below his reservation price—but probably not much below and not for long. Agents 7 and 8 also have a reservation price of $0.70, so if either one of them notices that agent 9 is snapping up certificates at (say) $0.60, one or both are likely to begin to bid (say) $0.62 or $0.65 in an attempt to acquire the certificates. This process may be expected to raise the transacted price to $0.70 in fairly short order (recall the 1-cent trading commission) as long as at least two of the type III traders actively attempt to buy up "cheap" certificates. Evidently $0.70, the highest reservation price,[8] is the equilibrium price for period C, and the corresponding allocation involves all certificates being held by the type III agents. Presumably the most aggressive bidder among them gets the lion's share.

---

[7] Inconclusive inside information (a series of "clues" as to the true state) has been examined experimentally in Experiment 1 of PS. They find that convergence to the rational expectations equilibrium is much slower in this case compared to experiments with conclusive inside information. Plott and Sunder (1983) also implement inconclusive inside information (no trader knows the true state, but their pooled information set is conclusive).

Now consider trading in the prior period B. At this point our agent 9 knows a certificate will yield him at least $0.30 (current dividend) + $0.70 (period C dividend) = $1.00, and possibly more if he were able to resell in the current period at a price above $1.00 or in per[illegible] a price above $0.70. We are unable to determine his reserva[illegible] in this period without knowing his attitude to risk and the p[illegible] ties he assigns to these resale possibilities, but we do know [illegible] bounded below by his "security level" of $1.00. Agent 1 is in [illegible] interesting position: he has a lower security level ($0.45 + [illegible] $0.90), but if he has seen this market operate for several Marke[illegible] he may become quite confident of his ability to resell in period C a[illegible] $0 70. In this case his reservation price will be approximately $0.45 (current dividend) + $0.70 (resale value in period C) = $1 15, so he is likely, at least in later Market Years, to outbid agent 9 and the other type III agents. Once again the price should be bid up to the highest reservation price among the agents of type I if the others (agents 2 and 3) also come to realize (either independently or through noticing agent 1's "windfall profits") that certificates might be worth more than $1.00 to them in period B. Consequently, the period B price should settle somewhere between $1.00 (the highest security level) and $1.15 (highest "rational" reservation price), with the price rising toward the higher value as agents accumulate experience over the years. All *certificates should be held by type I traders purchasing at prices in* excess of $1.00.

---

[8] More specifically, the second highest reservation price among the nine traders, which in the current instance happens to equal the first and third highest since there are three agents of type III. See Vickrey (1961) for an explanation of the familiar pricing result that the winning bid is the second highest reservation price. Friedman (1982) and Friedman and Harrison (1984) formalize the logic and Bayesian computation of these reservation prices for models of the price formation process in experimental asset markets.

Similar considerations apply to period A trading. The highest security level belongs to type I traders ($0.45 + $0.45 + $0.45 = $1.35), but if type II traders begin to anticipate the ability to resell in period B at a price approaching $1.15 their reservation price would be near $0.70 + $1.15 = $1.85. Over the Market Years, learning may allow the closing period A price to rise from $1.35 to nearly this value.

Suppose now that the data in table 3 refer to an experiment in which futures trading is permitted, and consider the strategic possibilities available to agent 1 in period B. He personally would be willing to bid at most $0.45 for futures contracts (i.e., period C delivery of certificates), but for reasons discussed above it seems likely that agents of type III will bid up the price of this contract to $0.70. With a currently quoted futures price of $0.70, agent 1 now has a security level for spot transactions of $1.15, not $0.90 as before: for each certificate he acquires he can still earn his current period dividend of $0.45 while "locking in" an additional $0.70 return by selling a futures contract. He need not guess, as was the case in the spot-only market, what its subsequent resale value might be. Hence a period B spot price of $1.15 and a futures price of $0.70 can arise from information *directly available to the agents. Note that spot trading in period C should be redundant*, given the operation of the futures market in prior periods.

The strategic opportunities in period A are quite rich. The opportunities discussed above still obtain, but the futures price provides additional information. If the futures price for period C delivery of certificates converges to $0.70 in period A, then agents of type II will have a security level of $1.70 ($0.70 current dividends + $0.30 period B dividends + $0.70 proceeds from the futures transaction). Hence we may expect that these agents (together with type I agents) will more quickly bid up spot prices when the futures market is active.

So far we have only dealt with the market uncertainty design of Experiments 1 and 2. To see the effect of event uncertainty, suppose the parameters of table 2 apply to state X, and for simplicity assume that state Y dividends are each $0.10 higher than those for state X. Inasmuch as all traders know whether state X or state Y obtains before period B opens, there are no new considerations introduced into the previous discussion of period C and period B trading. However, consider the period A problem facing our favorite agent, 9, in Experiment 3 (spot trading only). If he knew that state X would eventuate his security level would be \$0.10 + \$0.30 + \$0.70 = \$1.10, but it would be \$0.20 + \$0.40 + \$0.80 = \$1.40 in state Y. Given that his priors have been induced as 0.5/0.5 over the two states, we may suppose that his "effective" security level is \$1.25 (or a little less with risk aversion). If trader 9 and his fellow traders initially consider only "buy and hold" strategies (i.e., they do not speculate on being able to resell or repurchase at favorable prices in subsequent periods), the arguments advanced above lead us to predict a market price equal to the highest such "effective" security level, which still belongs to type I agents; that is, \$1.35(0.5) + \$1.65(0.5) = \$1.50. On the other hand, once traders gain experience with trading opportunities in periods B and C in *both* states of nature, they may be willing to consider expected trading opportunities in forming reservation prices, in which case period A prices will generally tend to rise in later Market Years. How far? Given a maximum reservation price of \$1.85 in state X (as argued above) and \$2.15 in state Y, an upper bound of \$2.00 = \$1.85(0.5) + \$2.15(0.5) is defined.

Similar considerations apply when futures markets exist, as in Experiment 4. Given our hypothetical parameters, we would expect to see futures trading in period A yield a price for period C certificates of at least \$0.70 (the maximum security level in the least favor state, X) and probably closer to \$0.75 = \$0.70(0.5) + \$0.80(0 "effective" security level. Once the futures price has stabiliz prices in period A should reflect the higher security level an converge more rapidly to the upper bound of \$2.00. It is inte to observe that fairly complicated trading plans are often requ support these higher period A spot prices. In the current hyp ical example, type II traders who purchase spot certificates at a period A price near \$2.00 probably also hedge (sell futures) at the same time and plan in period B to sell spot while offsetting their futures positions. In the actual experiments, contingencies based on the state of nature are important, and efficient plans can become even more complex.

We close this discussion with a few observations regarding insider information. Once again, only period A trading is affected; to the extent that the insiders' information leaks out, the "event uncertainty" becomes irrelevant. Thus it is not clear a priori whether the state-specific or "effective" reservation prices are more appropriate, but we should often be able to determine from transactions data whether or not a leak occurred.◆

TABLE 3

SUMMARY OF EQUILIBRIUM CONCEPTS

Weaker Efficiency		Stronger Efficiency	
Experiment 1. Market Uncertainty with No Futures Market			
$p(\mathrm{A}\vert\mathrm{PI}) = \max\limits_i \sum_{k=1}^{3} d_i^k = \$1.35$	[I]	$p(\mathrm{A}\vert\mathrm{PF}) = \sum_{k=1}^{3} \max\limits_i d_i^k = \$1.85$	[II]
$p(\mathrm{B}\vert\mathrm{PI}) = \max\limits_i \sum_{k=2}^{3} d_i^k = \$1.00$	[III]	$p(\mathrm{B}\vert\mathrm{PF}) = \sum_{k=2}^{3} \max\limits_i d_i^k = \$1.15$	[I]
$p(\mathrm{C}\vert\mathrm{PI}) = \max\limits_i d_i^3 = \$0.70$	[III]	$p(\mathrm{C}\vert\mathrm{PF}) = \max\limits_i d_i^3 = \$0.70$	[III]
Experiment 2. Market Uncertainty with a Futures Market			
$p(\mathrm{A}\vert\mathrm{PI}) = p(\mathrm{C}\vert\mathrm{PI}) + \max\limits_i (d_i^1 + d_i^2) = \$1.75$	[I]	$p(\mathrm{A}\vert\mathrm{PF}) = \sum_{k=1}^{3} \max\limits_i d_i^k = \$2.00$	[I]
$p(\mathrm{B}\vert\mathrm{PI}) = p(\mathrm{C}\vert\mathrm{PI}) + \max\limits_i d_i^2 = \$1.25$	[II]	$p(\mathrm{B}\vert\mathrm{PF}) = \sum_{k=2}^{3} \max\limits_i d_i^k = \$1.25$	[II]
$p(\mathrm{C}\vert\mathrm{PI}) = \max\limits_i d_i^3 = \$0.80$	[III]	$p(\mathrm{C}\vert\mathrm{PF}) = \max\limits_i d_i^3 = \$0.80$	[III]
Experiment 3. Event Uncertainty with No Futures Market			
$p(\mathrm{A}\vert\mathrm{EPI}) = \max\limits_i \left[\sum_{k=1}^{3} \sum_{z\in Z} \pi(z) d_i^k(z)\right] = \$1.44$	[III]	$p(\mathrm{A}\vert\mathrm{URE}) = \max\limits_i \left\{\sum_{z\in Z} \pi(z)[d_i^1(z) + p(\mathrm{B}\vert\mathrm{PF}, z)]\right\} = \$1.85$	[I]
$p(\mathrm{B}\vert\mathrm{PI}, z) = \max\limits_i \sum_{k=2}^{3} d_i^k(z) = \$1.23 \begin{cases} z = \mathrm{X} \\ z = \mathrm{Y} \end{cases}$	[III] [I]	$p(\mathrm{B}\vert\mathrm{PF}, z) = \max\limits_i d_i^2(z) + p(\mathrm{C}\vert\mathrm{PF}, z) = \$1.40$ (X or Y)	[II]
$p(\mathrm{C}\vert\mathrm{PI}, z) = \max\limits_i d_i^3(z) = \$0.70 \begin{cases} z = \mathrm{X} \\ z = \mathrm{Y} \end{cases}$	[III] [I]	$p(\mathrm{C}\vert\mathrm{PF}, z) = \max\limits_i d_i^3(z) = \$0.70 \begin{cases} z = \mathrm{X} \\ z = \mathrm{Y} \end{cases}$	[III] [I]

Experiment 4 Event Uncertainty with a Futures Market

$p(A|EPI) = p(C|EPI) + \max_i \left[\sum_{k=1}^{2} \sum_{z\in Z} \pi(z) d_i^k(z)\right] = \$1.85$ [II]

$p(B|PI, z) = \max_i \sum_{k=2}^{3} d_i^k(z) = \$1.60 \begin{cases} z = X \\ z = Y \end{cases}$ [III]
[I]

$p(C|EPI) = \max_i \left[\sum_{z\in Z} \pi(z) d_i^3(z)\right] = \$0.85$ [III]

$p(C|PI, z) = \max_i d_i^3(z) = \$1.00 \begin{cases} z = X \\ z = Y \end{cases}$ [III]
[I]

$p(A|URE) = \max_i \left\{\sum_{z\in Z} \pi(z)[d_i^1(z) + p(B|PF, z)]\right\} = \$2.30$ [I]

$p(B|PF, z) = \max_i d_i^2(z) + p(C|PF, z) = \$1.80 \begin{cases} z = X \\ z = Y \end{cases}$ [II]
[II]

$p(C|URE) = \sum_{z\in Z} \pi(z) p(C|PF, z) = \$1.00$ [I, II, or III]

$p(C|PF, z) = \max_i d_i^3(z) = \$1.00 \begin{cases} z = X \\ z = Y \end{cases}$ [III]
[I]

Experiment 5 Insider Uncertainty with No Futures Market

With leakage: $p(A|PI, z)$ = $1.35 (X) or $1.70 (Y)
[I] [III]

No leakage: $p(A|EPI)$ = $1.25 [III]

$p(B|PI, z)$ = $0.85 (X) or $1.30 (Y)
[I or II] [III]

$p(C|PI, z)$ = $0.50 (X) or $0.80 (Y)
[I] [III]

With leakage $p(A|PF, z)$ = $1.45 (X) or $1.80 (Y)
[I] [III]

No leakage $p(A|URE)$ = $1.575 [I]

$p(B|PF, z)$ = $0.95 (X) or $1.40 (Y)
[II] [II]

$p(C|PF, z)$ = $0.50 (X) or $0.80 (Y)
[I] [III]

Experiment 6 Insider Uncertainty with a Futures Market

With leakage: $p(A|PI, z)$ = $1.15 (X) or $1.55 (Y)
[I] [III]

No leakage $p(A|EPI)$ = $1.225 [III]

$p(B|PI, z)$ = $0.75 (X) or $1.15 (Y)
[II] [III]

With leakage: $p(C|PI, z)$ = $0.45 (X
[II]

No leakage $p(C|EPI)$ = $0.50

With leakage $p(A|PF, z)$ = $1.30 (X) or $1.80 (Y)
[I] [II]

No leakage $p(A|URE)$ = $1.45 [I]

$p(B|PF, z)$ = $0.85 (X) or $1.30 (Y)
[III] [II]

With leakage $p(C|PF, z)$ = $0.45 (X) or $0.70 (Y)
[II] [III]

No leakage $p(C|URE)$ = $0.575 [I, II, or III]

NOTE.—Type of trader (I, II, or III) is indicated in square brackets [after or below the price. Formulae are omitted for Experiments 5 and 6

B. A Theoretical Model

A very brief sketch of our theoretical model of experimental markets is required to define and justify our predicted market outcomes. A more complete account can be found in Appendix A.

Agents, or traders, indexed $i = 1, \ldots, 9$, are classified into types (I, II, or III) according to (i) their endowments $x_i(0)$ of cash and period-dated assets, with periods indexed $k = 1, 2, 3$, and (ii) their preferences, determined by specific dividends per certificate, $d_i^k(z)$, where $z = X$ or Y denotes the state of nature.[9] We assume each agent i will try to transact so as to acquire final holdings $x_i(T)$ that maximize end-of-year wealth,

$$W_i(T, z) = d_i(z) \cdot x_i(T) \equiv x_i^0(T) + \sum_{k=1}^{3} d_i^k(z)x_i^k(T) \equiv \text{cash} + \text{dividends.}$$

Of course, trading typically occurs under incomplete and possibly asymmetric information regarding the state of nature, z, and the trading opportunities contingent on z. The strongest notion of informational efficiency is based on the idea that the market has somehow disseminated perfect information regarding both. We refer to this case as perfect foresight (PF); the equilibrium is then the ordinary (complete markets, perfect information) competitive equilibrium in which all certificates are held by traders of the highest dividend type in each period and prices reflect these dividends. That is, given that the state of nature is (or will be revealed as) z, the PF prices are

$$p(\mathrm{A}|\mathrm{PF}, z) = \sum_{k=1}^{3} \max_i d_i^k(z) \text{ in period A,}$$

$$p(\mathrm{B}|\mathrm{PF}, z) = \sum_{k=2}^{3} \max_i d_i^k(z) \text{ in period B,}$$

$$p(\mathrm{C}|\mathrm{PF}, z) = \max_i d_i^3(z) \text{ in period C.}$$

The corresponding allocations x_i^k are zero unless i belongs to the type for which $d_i^k(z)$ is maximal,[10] $k = 1$ for period A, $k = 2$ for period B, and $k = 3$ for period C. For Experiments 1, 2, 5, and 6, this PF equilibrium corresponds directly to the popular notion of strong-form efficient prices.

---

[9] We make the convention that $z = X$ with certainty in Experiments 1 and 2, and often we will drop the index z when discussing these experiments.

[10] Inasmuch as the distribution of certificates is indeterminate among those of the given (maximal) type, the equilibrium is not unique. Note, however, that equilibrium *prices* are unique because the induced preferences are additively separable—cf. Arrow and Hahn (1971, chap. 9).

In Experiments 3 and 4 no information regarding the state of the world is available during period A trading. Hence the strongest available concept of informational efficiency must be based only on the contingent trading opportunities together with the induced prior probabilities. We refer to this case as *uninformed rational expectations* (URE): if it is assumed that all agents are risk neutral, know all the contingent (PF) prices, and share the probabilities $\pi(z)$ over states of the world $z \in Z = \{X, Y\}$, one can show (see App. A) that the equilibrium spot price in period A is

$$p(A|URE) = \max_{i} \sum_{z \in Z} \pi(z)[d_i^1(z) + p(B|PF, z)].$$

Likewise,

$$p(C|URE) = \sum_{z \in Z} \pi(z) p(C|PF, z)$$

is the equilibrium futures price in period A under the same circumstances. As usual, the corresponding period A allocations for spot certificates are zero for nonmaximal i and indeterminate among i in the maximal type. Futures allocations in period A are completely indeterminate; indeed, under current assumptions the period A futures market is redundant, given the opportunity in period B to readjust positions before period C dividends are distributed.

It is interesting to note that in Experiments 5 and 6 both URE and PF equilibria are relevant, with URE in this case corresponding to the notion of semistrong informational efficiency (prices reflect all public, but not insider, information).

In the numerical example of the previous subsection, PF or U prices were regarded as upper bounds that might be approache traders acquired information of the second type. Since traders pos no reliable information of this sort at the beginning of each exp ment, it is useful to define another benchmark equilibrium co sponding to weaker concepts of information efficiency. We refer t the case in which traders know the state of the world and their own personal dividend profiles but not their trading opportunities in subsequent periods as private information (PI). There is an incomplete-markets, incomplete-information competitive equilibrium (in which agents do *not* use equilibrium prices as a supplementary source of information) under these circumstances whose contingent prices are[11]

---

[11] As usual, the corresponding allocations are zero except for i of the type for which the indicated maximum is achieved. This will often be taken for granted in later discussions.

$$p(\mathrm{A}|\mathrm{PI}, z) = \max_i \sum_{k=1}^{3} d_i^k(z),$$

$$p(\mathrm{B}|\mathrm{PI}, z) = \max_i \sum_{k=2}^{3} d_i^k(z),$$

$$p(\mathrm{C}|\mathrm{PI}, z) = \max_i d_i^3(z).$$

Two possible modifications to the PI prices for period A should be mentioned briefly. First, if futures markets operate, then markets are not quite so incomplete and a slightly stronger notion (still denoted PI) is appropriate:

$$p(\mathrm{A}|\mathrm{PI}, z) = p(\mathrm{C}|\mathrm{PI}, z) + \max_i \sum_{k=1}^{2} d_i^k(z),$$

where $p(\mathrm{C}|\mathrm{PI}, z) = \max_i d_i^3(z)$ is, of course, also the equilibrium contingent *futures* price in period A. Note that the futures market is not redundant under PI assumptions.

Second, if agents do not know z in period A but share firm priors $\pi(z)$ we may define the expected private information (EPI) equilibrium price as

$$p(\mathrm{A}|\mathrm{EPI}) = \max_i \sum_{k=1}^{3} \sum_{z \in Z} \pi(z) d_i^k(z).$$

An interpretation of EPI based on "effective security levels" was offered in the previous subsection.

For convenient reference we list and compute the applicable equilibria for each experiment in table 3.

The relevance of all these equilibrium concepts may not be immediately apparent to the reader. Our markets are not populated by large numbers of price-taking agents, nor is Walrasian "recontracting" allowed, so the conventional theoretical basis for competitive equilibrium is absent. Instead, our double-auction institution creates a very complex, continuous-time extensive-form game of incomplete information for our set of nine agents. Friedman (1982) shows how some apparently mild consistency assumptions can lead to a tractable analysis of single-period games of this type, and in Appendix A we extend that analysis to 3-period asset markets. The first main result is that PI *final* allocations and prices can be supported by a certain noncooperative equilibrium in strategies contingent on very limited information. After noting that agents have more public information in later Market Years since they have been able to observe market prices in all periods, we show that PF (or URE) equilibria can be supported as noncooperative equilibria in strategies based on information that is eventually made public by the double-auction process.

C. Predictions from the Theory

Convergence to stronger-form informational efficiency.—Our first principal prediction is that prices and allocations in each experiment more nearly approach the stronger-form equilibrium values in later Market Years. The reason, of course, is that traders acquire better information as to (state-contingent) trading opportunities through their actual trades, and this information becomes reflected in the prices. Although theorists may differ on the details of the process, its existence is generally uncontroversial, and virtually all experimental double-auction markets display a powerful tendency to approach full efficiency under stationary repetition.[12]

This first prediction does involve one controversial point in connection with our Experiments 5 and 6. Is the semistrong form of informational efficiency (URE) or the strong form (PF) the better predictor of behavior in later Market Years? Given that we have only one insider of each type competition may not force the inside information into the open; indeed, we know of no serious theoretical argument for PF efficiency for this case.[13] On the other hand, those economists with great faith in the "wisdom of the market" would probably predict this strongest-form efficiency even without the theory.

The informational role of futures markets.—Our second predicti that the presence of a futures market speeds the approach to st er-form equilibrium and reduces the volatility of transacted prices. The theoretical justification is clear: markets are more n complete and more information is available when futures marke open. To drive this point home, consider the minimum time requ for PF (or URE) outcomes to emerge from an ideal learning proces In the spot-only, no-event-uncertainty design of Experiment 1, traders should ideally observe $p(C|PF) = p(C|PI) = \max_i d_i^3$ at the end of the first Market Year (see App. A). If they all jump to the conclusion that $p(C|PF)$ will always obtain in period C of subsequent Market Years, believe that other traders have jumped to the same conclusion, and trade accordingly, then they will be correct. Furthermore, in this case the same sort of argument shows that period B prices will close at

---

[12] Harrison, Smith, and Williams (1983) present experimental evidence for perishables that convergence to equilibrium in double auctions is sluggish in the absence of market stationarity and that deviations from equilibrium are often sustained and severe. Friedman and Harrison (1984) present evidence for single-period asset markets that convergence to equilibrium is more rapid in the absence of stationarity

[13] Appendix A sketches an argument for at least URE efficiency.

$p(\mathrm{B}|\mathrm{PF})$ in the following repetition (year 2). Similarly, closing prices in period A of year 3 can reach $p(\mathrm{A}|\mathrm{PF})$ if agents count on trading at $p(\mathrm{B}|\mathrm{PF})$ in the subsequent period. Hence the PF outcome is possible in year 4 for Experiment 1.

The existence of futures markets can shorten this learning process by a Year, since $p(\mathrm{B}|\mathrm{PI}) = p(\mathrm{B}|\mathrm{PF})$ if futures are traded in period B. In the case of event uncertainty one must observe $p(\mathrm{C}|\mathrm{PF}, z)$ and $p(\mathrm{B}|\mathrm{PF}, z)$ for each $z \in Z$ before $p(\mathrm{A}|\mathrm{URE})$ can emerge, so the minimum learning time is $2(\#Z) + 1$ years if there is no futures market (Experiment 3) and $\#Z + 1$ if there is a futures market (Experiment 4). The presence of insiders would not lessen the theoretical minimum learning time, although it could lessen (or increase, since noninsiders may be more cautious) the actual learning time.

As noted in our literature survey, the prediction that futures markets are stabilizing and speed convergence to informationally efficient equilibria is not without controversy. The contrary argument by Svensson (1976) for destabilization employs the assumption that agents may unanimously hold incorrect point expectations regarding subsequent prices, and therefore its relevance to our experimental markets is questionable. However, the popular version of the contrary view, that speculation in futures markets introduces "noise" into the price system, may have a counterpart in our experiments. Our traders face a more complex environment when futures trading is permitted: they must keep track of two prices instead of one in each trading period, the total trading time in each Market Year is less because there are two trading periods instead of three, and more complicated buying/selling decisions may be required because there are four "markets" (spot and futures in periods A and B) instead of three. Hence they are more likely to suffer from "information overload" and confusion, and market efficiency could thereby be impaired (see Grether and Wilde 1982).

IV. Experimental Results

Our raw data provide an embarrassment of riches: for spot and (where applicable) futures, in each period of each Market Year in each experiment we have a long sequence of bid and ask prices, accepted or unaccepted, identified by trader. Each such sequence typically contains 10–50 entries. Such detail is critical when studying the real-time process of learning behavior and price formation as in Easley and Ledyard (1983), Harrison, Smith, and Williams (1983), and Friedman and Harrison (1984), but it tends to obscure our present concerns. Therefore we reduce the data by stages. First, in the next subsection we take a brief look at the time series of *transacted* (accepted) prices in each experiment. This overview provides a qualitative assessment of our principal predictions and a sound basis for subsequent statistical analysis. A more detailed "market commentary" that ferrets out and attempts to explain all anomalies is provided in Appendix B.

The statistical analysis in the final subsection is performed on a data set that further reduces each sequence to a single observation. We then have only about 90 observations (6 experiments × 5 years × 3 markets),[14] but our orthogonal experimental design allows considerable mileage from this smaller data set. For instance, a given Experiment 3 observation can be used in an evaluation of (*a*) the effect of event uncertainty, in a comparison with Experiment 1; (*b*) the effect of futures markets, in a comparison with Experiment 4; and (*c*) the effect of "insiders," in a comparison with Experiment 5. Our statistical analysis will not be exhaustive, however; in the interest of brevity we will only present evidence bearing on our principal predictions.

A. Overview

Figures 1–6 chart the time series of transacted prices in each market experiment. The relevant theoretical equilibria for each period are indicated with dashed lines, and summary statistics of transacted prices are shown at the bottom. Where applicable an X refers to spot prices and a 0 to futures prices. Table 4 presents the observed period-end spot allocations, with the main theoretical predictions indic by asterisks. (Period C and futures allocations are listed in table App. C.)

Turning first to figure 1, we see that in Experiment 1 (no uncertainty, no futures market), period A and B prices both to rise from values near PI in the first Market Year to near PF by the end of year 5. Period C prices seemed to settle near their PI PI value by the end of year 2. Allocations also converged to the PF predictions at about the same rate, as indicated in table 4. One could hardly hope for a clearer qualitative confirmation of our first prediction than this experiment provides.

Figure 2 shows that futures transactions in Experiment 2 all occurred at prices at or slightly below the equilibrium level of $0.80, even in early Market Years. Period A and B prices and allocations are also very near PF values by year 3.[15]

---

[14] In futures experiments there are four markets if one counts period A and period B trading separately for period C certificates.

[15] See FHS (pp. 143, 152, 156, 160) for a discussion of two anomalies (one induced by the experimenters, the other by a "deviant" subject) in Market Year 5 of this experiment, one can see that the effects are detectable but not large

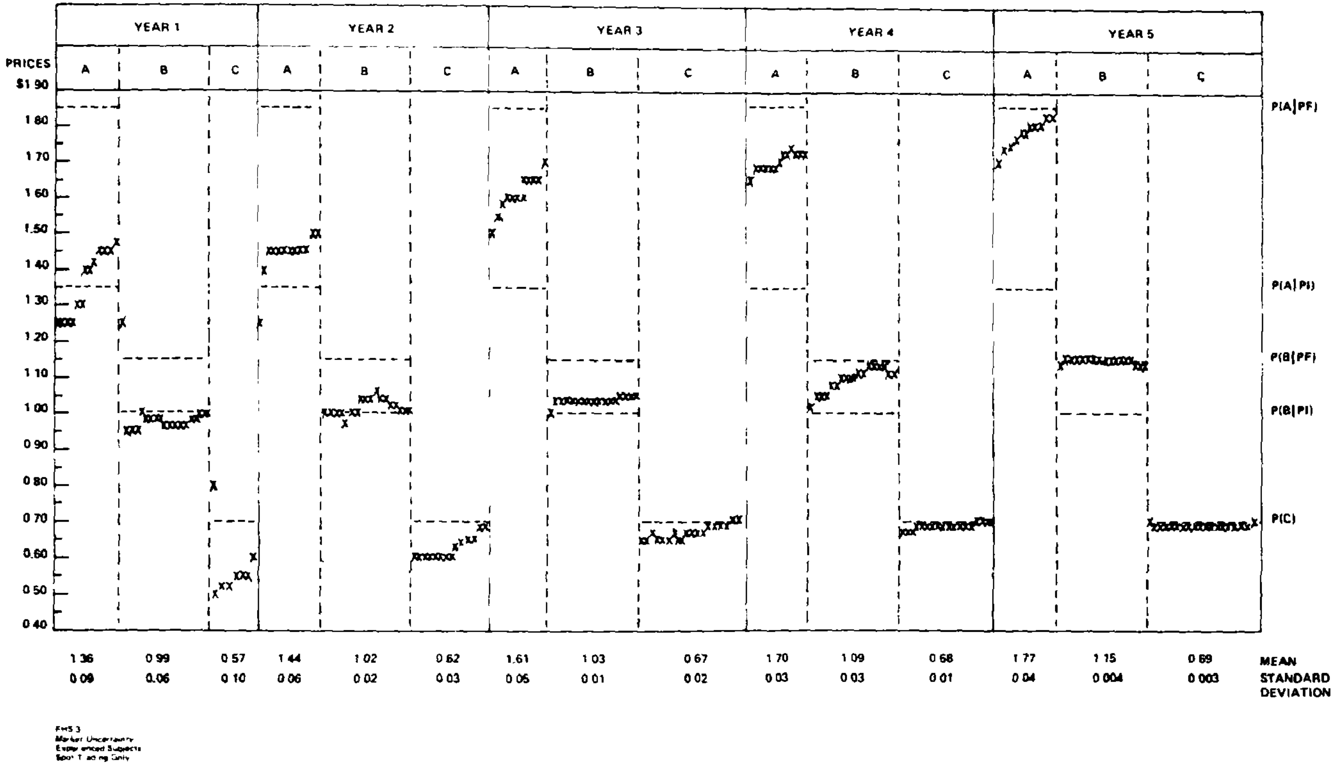
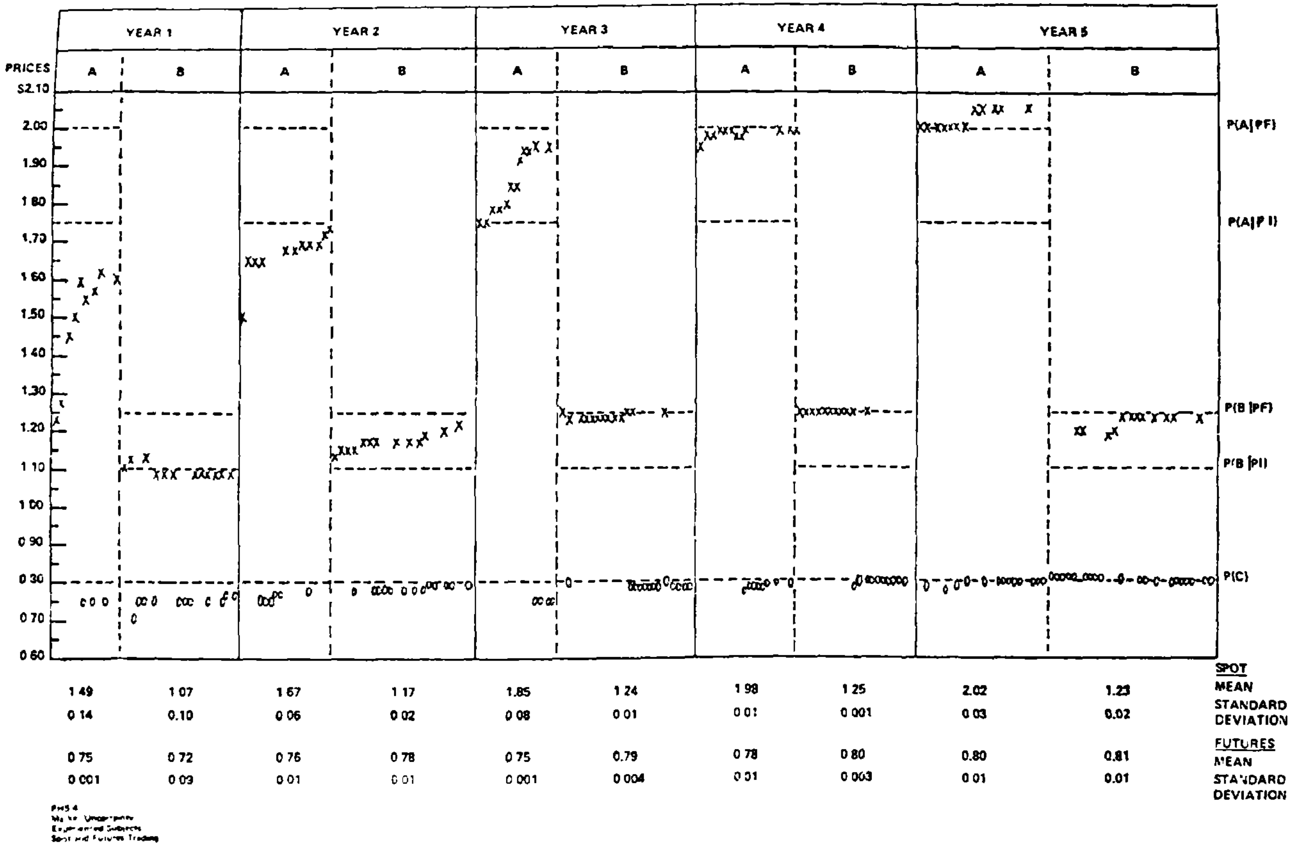


FIG. 1.—Observed price behavior, Experiment 1



Observed price behavior, Experiment 2

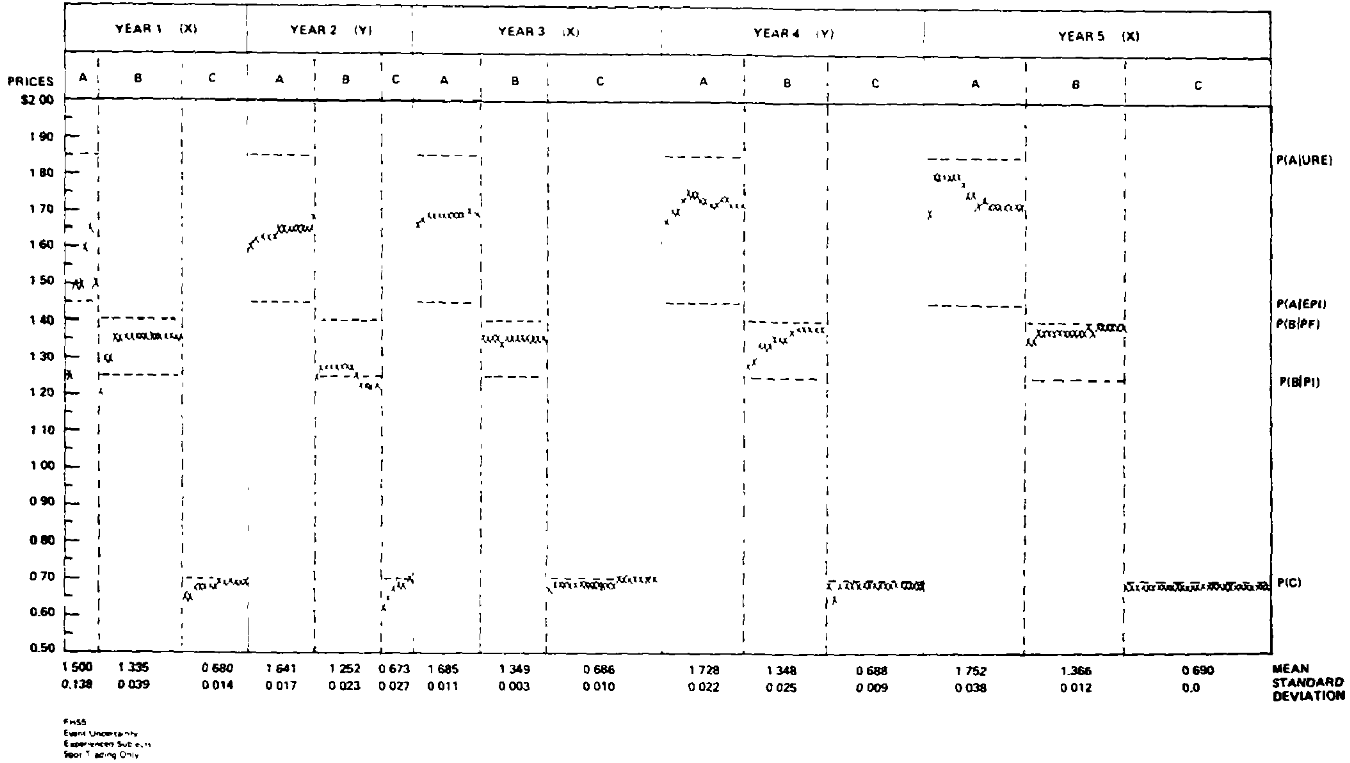
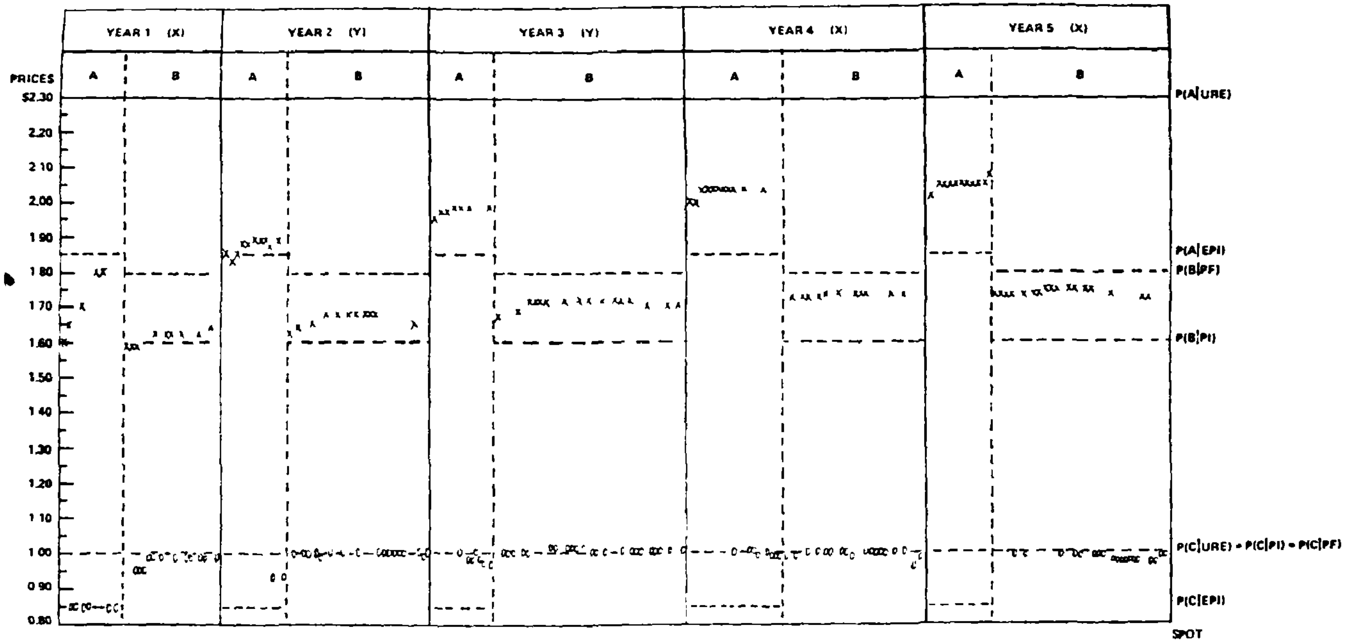


FIG 3 —Observed price behavior, Experiment 3



	YEAR 1 A	YEAR 1 B	YEAR 2 A	YEAR 2 B	YEAR 3 A	YEAR 3 B	YEAR 4 A	YEAR 4 B	YEAR 5 A	YEAR 5 B
SPOT MEAN	1 710	1 618	1 866	1 660	1 970	1 706	2 032	1 726	2 049	1 743
STANDARD DEVIATION	0 089	0 021	0 017	0 017	0 015	0 010	0 016	0 005	0 007	0 007
FUTURES MEAN	0 843	0 988	0 940	0 999	0 990	1 003	0 994	0 994		0 985
STANDARD DEVIATION	0 019	0 013	0 001	0 003	0 008	0 005	0 005	0 013		0 005

F ehavior, Experiment 4

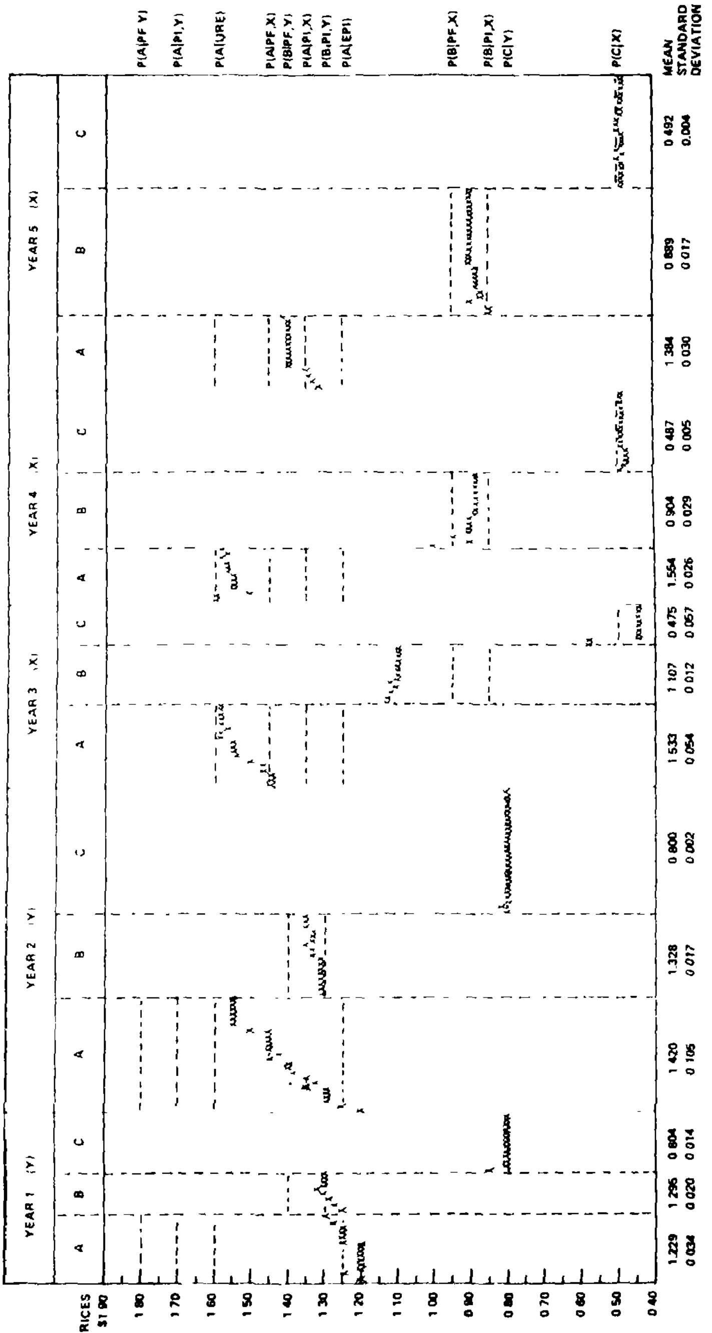
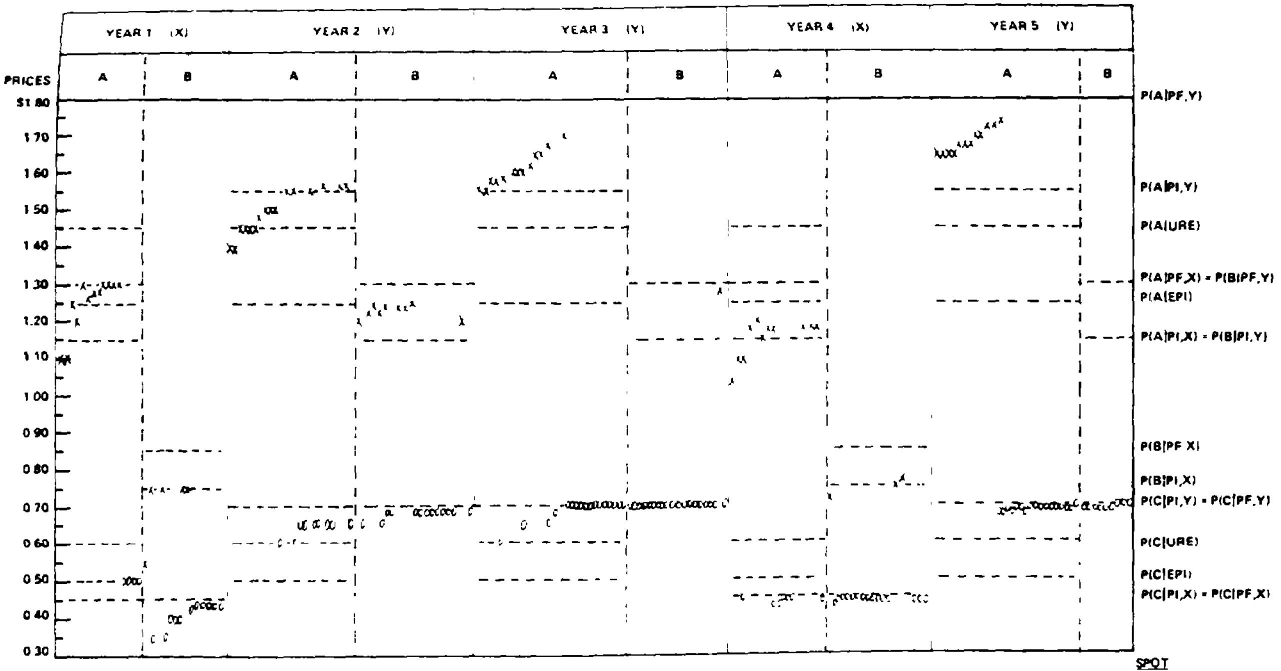


FIG. 5.—Observed price behavior, Experiment 5



	Year 1 A	Year 1 B	Year 2 A	Year 2 B	Year 3 A	Year 3 B	Year 4 A	Year 4 B	Year 5 A	Year 5 B
SPOT										
MEAN	1.248	0.710	1.498	1.223	1.608	1.280	1.155	0.740	1.686	
STANDARD DEVIATION	0.075	0.089	0.057	0.019	0.047		0.041	0.026	0.030	
FUTURES										
MEAN	0.500	0.404	0.631	0.678	0.687	0.700	0.439	0.439	0.688	0.694
STANDARD DEVIATION	0.0	0.031	0.026	[illegible]	0.028	0.0	0.007	0.003	0.005	0.005

Observed price behavior, Experiment 6

TABLE 4

SPOT CERTIFICATE ALLOCATIONS BY TRADER TYPE

			PERIOD A			PERIOD B		
EXPERIMENT	YEAR	STATE	I	II	III	I	II	III
1	1	...	4*	14**	0	3**	1	14*
	2	...	0*	18**	0	5**	2	11*
	3	...	0*	18**	0	16**	2	0*
	4	...	0*	18**	0	18**	0	0*
	5	...	0*	18**	0	18**	0	0*
2	1	...	13*/**	4	1	1	13*/**	4
	2	...	13*/**	5	0	0	18*/**	0
	3	...	13*/**	5	0	0	18*/**	0
	4	...	18*/**	0	0	7	11*/**	0
	5	...	14*/**	4	0	2	16*/**	0
3	1	X	5**	3	10*	0	17**	1*
	2	Y	14**	0	4*	13*	5**	0
	3	X	10**	0	8*	0	14**	4*
	4	Y	11**	0	7*	0*	15**	3
	5	X	18**	0	0*	0	18**	0*
4	1	X	6**	5*	7	3	13**	2*
	2	Y	11**	5*	2	5*	12**	1
	3	Y	9**	0*	9	2*	16**	0
	4	X	10**	0*	8	7	11**	0*
	5	X	8**	0*	10	1	17**	0*
5	1	Y	6	0	12*/**	0	1**	17*
	2	Y	0	0	18*/**	0	18**	0*
	3	X	14**	0	4*	8*	6*/**	4
	4	X	7**	0	11*	2*	14*/**	2
	5	X	3**	13	2*	2*	16*/**	0
6	1	X	6**	5	7*	7	8*	3**
	2	Y	0	10**	8*	1	17**	0*
	3	Y	0	18**	0*	0	18**	0*
	4	X	13**	2	3*	16	1*	1**
	5	Y	0	18**	0*	0	18**	0*

NOTE.—A single asterisk indicates the weaker form prediction (PE, PE, EPE, EPE, EPE, and EPE, respectively, in period A for each experiment; PE in period B in all experiments) and a double asterisk indicates the stronger form predictions ([illegible] and PE, respectively, in period A for each experiment; PE in period B).

Recall that two states of the world were possible in Experiment 3, but by choice of parameters period B and C equilibrium prices were independent of realized state, while period A dividends and most equilibrium allocations were state contingent. With this in mind, it is not surprising to see fairly rapid convergence of period C prices and rather steady year-to-year movement of period B prices from PI to PF values, despite the yearly change in the "state of nature." Period A trading, thin and erratic with average prices near the EPI value in year 1, becomes fairly steady by years 4 and 5 with allocations precisely at their URE values and prices a little ($0.05–$0.15) below the URE level. One could attribute the low prices to risk aversion, incomplete learning, or both.

Experiment 4 contains both expected and unexpected behavior. Futures prices behaved much as one might expect, rising from their EPI value of $0.85 in period A of year 1 to their URE and PF value of $1.00 by period B of year 2. Futures positions on the whole followed suit. Spot prices, on the other hand, were quite sluggish: average period B prices rose each year but did not complete the transition from PI to PF levels, while period A prices only made about half of the transition from EPI to URE levels. As a consequence of these low period A prices, it was profitable for type III traders to purchase spot even though they could not count on an X realization of "nature." Evidently, type I traders were slow to recognize significant expected profit opportunities arising from their period A spot purchases.

A glance at figure 5 discloses that the presence of insiders is not the only complication—parameters were chosen so that virtually all equilibrium prices and allocations were state contingent. Period C prices and allocations tracked the equilibrium rather well with a minor exception in year 3, the first X year. There is a strong hysteresis effect in period B prices of the same year that dominates the otherwise even rise from PI to PF levels.[16] The same effect is present in period A prices, which otherwise seem to move from EPI toward URE levels. This last interpretation, suggesting that inside information was *not* leaked, is supported by the fact that in period A of the last two years the majority of certificates were *not* held by type I agents, as predicted in either leakage equilibrium (PI or PF).

A careful study of figure 6, on the other hand, leads to the strong suspicion that inside information leaked out to the public rather quickly in Experiment 6. Futures prices in period A consistently approached the informed (PI = PF) equilibrium values of $0.45 in X years and $0.70 in Y years, providing a good signal to noninsiders.

---

[16] We refer here to the tendency of the market to respond sluggishly to a change in the state of nature, so that previously established behavior prevails for a while. See Harrison et al. (1983) for a discussion of such phenomena.

Despite the fact that the state of nature changed three times in this experiment, there is much less hysteresis in period A and B spot prices than in Experiment 5, in which the state only changed once. In general the period A spot price and allocation trends seem much more consistent with PF than the alternative equilibria, evidence in support of strong-form informational efficiency.

B. Statistical Analysis

Convergence to stronger-form informational efficiency.—We now evaluate the prediction that our asset markets approach the stronger forms of informational efficiency by examining prices, allocations, and profits. The traditional approach in the experimental literature has been to use a t-test on the sequence of transacted prices in a given trading period to see if the mean "differs significantly" from the theoretical equilibrium price. Once one recognizes that the transacted prices are not at all analogous to i.i.d. samples from a fixed population,[17] the meaning of such tests becomes quite obscure. Fortunately, the very multiplicity of theoretical equilibria (for prices, allocations, and profits) provides an attractive alternative: we can see which equilibrium concept gives us the best forecast. We are interested in the improvement or deterioration of the forecast performance of each concept, its overall accuracy, and its terminal accuracy in the final year of trading.

An appropriate measure of price convergence is the root mean squared deviation (RMSD) of transacted from forecast (equilibrium) price. Our prediction then is that RMSDs will decline (i.e., forecasts improve) over Market Years in both periods A and B for stronger-form equilibrium prices (PF or URE) and that RMSDs will rise (i.e., forecasts deteriorate) for weaker-form equilibrium prices (PI or EPI). We also predict that the strong-form forecasts will be more accurate in the final year of trading and perhaps more accurate overall.

The data in table 5 strongly support these predictions. In period A of Experiment 1, for instance, the PI forecast error increases monotonically from \$0.085 in year 1 to \$0.427 in year 5, while the PF forecast error declines monotonically from \$0.501 to \$0.082. Similarly, PF forecast errors for period B decline steadily from \$0.170 to \$0.005 while PI errors, after dropping from \$0.062 in year 1 to \$0.022 in year 2, increase steadily to \$0.148 in year 5. The data for Experiment 2 tell a very similar story. Likewise, in Experiments 3 and

---

[17] Bear in mind that each transaction provides new public information to the traders, so the hypothetical "population of potential prices" changes continually *within* the trading period. Harrison et al. (1983) and Friedman and Harrison (1984) expand on this theme, which is also noted by FPP*a* (p. 549) when introducing their statistical tests.

4 there is a striking improvement in the (strong-form) URE forecasts and a general deterioration in the (weak-form) EPI forecasts for period A over successive Market Years of either realized state of nature.[18] The same behavior is evident with respect to the PF and PI forecasts for period B.

One can make similar observations with respect to Experiments 5 and 6, and for relevant period C prices, but the proliferation of numbers (e.g., four distinct forecasts for spot prices and three for futures prices in period A of Experiment 6) makes this sort of analysis increasingly awkward. For that reason we present statistics in table 6 that summarize the performance of each forecast for a given experiment. Each entry in the first ten columns of table 6 is the posterior probability that a particular forecast is more likely to improve than to deteriorate from one Market Year (of a given state of nature) to the next.[19] Hence values above 0.5 for strong-form forecasts and below 0.5 for weak-form forecasts are evidence in favor of our prediction.

Table 6 confirms the conclusions drawn so far from table 5. For instance, the steady improvement noted in the PF forecasts for Experiment 1 yields posterior probabilities of .823 for both period A (col. 4) and period B (col. 6) and .895 for the pooled data (col. 10), while the corresponding entries for the PI forecasts (cols. 2, 5, and 8) yield probabilities for forecast improvement of only .177, .261, and .095, respectively. The results for Experiments 2–4 are generally similar, with pooled (period A + period B) probabilities of strong-form forecast improvement of .848, .848, and .895, respectively.

Recall that there are four competing equilibrium forecasts i
periments 5 and 6. In Experiment 5 there is a general improv
in the PI, URE, and PF forecasts for period A and no signi
improvement in either of the forecasts (PI and PF) for period B
pooled data for both periods therefore reflect the period A out
which allows us to dismiss only the EPI forecast (probability = 26
.5). It is interesting to note that in Experiment 6 the "leakage fore-

---

[18] Our comments here and the statistical tests presented below generally pool results for each experiment over the two states of nature. Looking separately at the sequences of years for each state would strengthen our conclusions with respect to our primary predictions but would reduce the sample size and lower the power of our tests.

[19] Each change in forecast performance in successive market years of a given state of nature in a given experiment is viewed as an outcome from a discrete binomial distribution, with a forecast improvement being viewed as a "success" and any nonimprovement a "failure." The diffuse prior distribution adopted for the probability of a "success," θ, was the Bayes-Laplace uniform prior on the unit interval. In table 6 we report the posterior probability that θ exceeds one-half. See Zellner (1971, pp. 38–40; p. 316, n. 37) for a complete discussion of these Bayesian procedures. A similar measure of forecast improvement is implied by a test used by Plott and Sunder (1982, p. 683), although their test was applied to individual price changes whereas ours is applied to changes in average forecast performance (RMSD) over each year.

TABLE 5

Root Mean Squared Deviations from Equilibrium Price Predictions

Experiment and Year	State	Period A				Period B		Period C			
		EPI	PI	URE	PF	PI	PF	EPI	PI	URE	PF
1:											
1	...	...	085	...	501	.062	.170	...	*	...	.158
2	...	...	.107	...	.415	.022	136	...	*	...	.082
3	...	...	266	...	.245	.037	.116	...	*	...	.036
4	...	...	.352	...	.152	.096	.066	...	*	...	.020
5	...	...	.427	...	.082	.148	.005	...	*	...	.010
2:											
1	...	...	.297	...	532	.155	155	...	*	...	050/.114
2	...	...	.104	...	342	.087	.087	...	*	...	.044/.018
3	...	...	.128	...	.164	010	010	...	*	...	.050/.009
4	...	...	.232	...	022	001	001	...	*	...	.019/.003
5	...	...	.272	...	.032	.031	.031	...	*	...	.007/.013
3:											
1	X	.139	...	372	...	.112	.075	...	*	...	.024
2	Y	.202	...	.209	...	.031	.150	...	*	...	.037
3	X	.245	...	165	...	119	051	...	*	...	.017
4	Y	289	...	124	...	.120	057	...	*	...	.015
5	X	.314	...	105	...	136	.036	...	*	...	.010

4											
1	X	.161	...	595	...	027	183	.019	*	.158	.018
2	Y	.023	...	434	...	062	141	090	*	.060	.003
3	Y	121		330	...	.107	094	.140	*	.013	.006
4	X	.182	...	269	...	.126	.074	144	*	.008	.014
5	X	199	...	251		.143	057	...	*	...	.016
5:											
1	Y	039	.473	348	572	019	.106	...	*	...	.014
2	Y	.199	.298	186	394	.032	074	...	*	...	.002
3	X	288	190	068	.098	257	.158	...	*	...	.059
4	X	305	205	033	.107	.061	054	...	*	...	.013
5	X	138	.045	193	.072	.043	.063		*	...	.009
6.											
1	X	.077	.121	215	089	089	161	000	*	075	.050/.055
2	Y	229	076	073	.307	.076	079	133	*	.061	073/.025
3	Y	.337	074	.165	.197	130	.020	.189	*	.115	030/ 0
4	X	127	040	.289	.151	.024	112	.062	*	.137	.013/.011
5	Y	412	139	238	118	...	...	188	*	.113	013/.007

NOTE.—An asterisk for the PI RMSD of period C prices indicates that the statistics are the same as the PF RMSD for that year (i.e., the PI price prediction equals the PF price prediction). A slash in the PF futures price statistics for Experiments 2 and 6 indicates the period A RMSD/period B RMSD. In Experiment 4 the FPI and URE RMSD for the futures prices reflect period A prices only.

TABLE 6

SIGNIFICANCE OF FORECAST IMPROVEMENT AND COMPARATIVE ACCURACY

MEASURE AND EXPERIMENT	FORECAST IMPROVEMENT										OVERALL FORECAST ACCURACY			TERMINAL FORECAST ACCURACY	
	Period A				Period B		Period A and Period B								
	EPI	PI	URE	PF	PI	PF	EPI	PI	URE	PF	Period A	Period B	Period A and Period B	Period A	Period B
Prices.															
1	..	177		.823	261	.823	.	.095	.	895	421	210	.264	5.2	29 6
2	.	261	.	.739	739	739	..	5		848	.5	5	.602	8 5	1.0
3	177	..	823	. .	261	5	095		848	.	790	.726	803	2.9/2.3	3.8/2.1
4	261	...	823	.	177	823	.095	.	895	..	0	345	.038	.8/.4	2.5/1.1
5	261	739	739	.739	.5	.5	287	713	.713	713	.421	.075	218	1 9/.7	.5/ 4
6	261	739	.5	.739	639	639	.383	798	.617	798	.726	343	534	8/3.5	.2/6.5
1, 3, 5	.095	287	904	.904	172	.828	037	121	.963	963	574	267	.491	.952	845
2, 4, 6	152	5	.848	.848	398	.878	157	417	.952	952	.186	.362	.283	.726	.790
All	035	328	.965	.965	.174	945	017	147	.994	994	.313	429	.371	.968	.938
Allocations.															
1		177	..	.823	177	823	.	.095	.	895	996	992	1 0	...	.
2	. .	261	.	.261	739	.739	.	.5	.	.5	.5	.5	5	..	...
3	.261		739	.	261	.739	095	.	.848	...	972	.992	.999	...	. .
4	177	.	5	.	.261	.5	095		5		.996	.996	1.0	...	..
5	.5	.261	.261	261	739	739	713	.5	5	5	726	.345	.602	.	...
6	5	.5	261	.739	261	739	287	.287	5	848	.986	982	.998	...	..
1, 3, 5	287	.095	.5	.713	172	943	328	058	807	942	1.0	996	998	.952	.972
2, 4, 6	152	287	287	.5	310	828	058	328	.807	807	990	996	.993	.970	.972
All	106	.058	328	.672	136	.976	.057	.057	.902	.968	.994	996	.997	.990	.998

casts" PI and PF dominate the corresponding "no leakage forecasts" EPI and URE, supporting the conclusion that leakage of inside information occurred in Experiment 6.

The alternative forecasts do not differ for period C *spot* prices from for futures prices arising in period B trading. The futures ("period C") prices arising in period A of Experiment 4 still follow the predicted pattern: according to table 5 the EPI forecasts deteriorated and the URE forecasts improved in each case. In Experiment 6 we have three alternatives to consider (EPI, URE, and PF). A ranking of their forecast improvement provides additional support for our prediction, with the PF forecast improving more than the URE forecast, which in turn improved more than the EPI forecast.[20]

Despite their impressive showing with respect to forecast improvement, the strong-form forecasts do not appear to dominate their weak-form counterparts with respect to *overall accuracy*. In columns 11–13 of table 6 we report the probability that the yearly RMSDs for the stronger-form forecast are on average lower than those for the corresponding weaker-form forecast.[21] These probabilities exceed .5, suggesting that the stronger-form forecast is more accurate overall for the pooled (period A + period B) sample in only three of the six experiments (2, 3, and 6).

This negative result is not surprising considering the relatively few repetitions of each state of nature in our experiments. Given the absence of any generally accepted models of learning and price adjustment, it has become common in the experimental literature to focus on comparative terminal accuracy. Of course, we have on observation for each state in each experiment when we examin tive forecast errors by this criterion, so we cannot report prob here. Instead we report the ratio of the weaker-form RMSD stronger-form RMSD in the last two columns of table 6 (top s only).[22] An entry of at least 1.0 is favorable for our prediction t strong-form forecasts are superior; we observe such entries in every case for the experiments in which the terminal year is the fifth (1 and 2) and in a majority of cases in which the terminal year is the second or third repetition (Experiments 3–5, which have X and Y states of nature).

---

[20] Table 6 does not show the probabilities of improvement associated with these remarks concerning futures prices in period A. In Experiment 4 they were .214 (EPI) and .786 (URE), and in Experiment 6 they were .261 (EPI), .5 (URE), and .739 (PF). Pooling the results for these two experiments we obtain probabilities .118 (EPI), .798 (URE), and .882 (URE in Experiment 4 and PF in Experiment 5, the strong-form equilibria).

[21] These probabilities are obtained from a (one-tail) nonparametric Mann-Whitney test for the "slippage problem." See Hoel (1971, pp. 310–18) for a formal discussion. We state the test outcomes in terms of the probability of the hypothesis that slippage has occurred in favor of the stronger-form equilibrium forecasts; the alternative hypothesis is that both samples arise from the same population. The weaker-form (stronger-form) period A equilibria used for this test in Experiments 1–6 are, respectively, PI (PF), PI (PF), EPI (URE), EPI (URE), EPI (PF), and EPI (PF). The relevant equilibria in period B are, of course, PI and PF in each experiment.

[22] The terminal X years and Y years for Experiments 3, 4, 5, and 6 are separated by a slash in table 6.

We may summarize our findings on price convergence by examining the statistics in table 6 obtained by pooling data across all experiments (line 9, marked "all"). Our prediction that strong-form equilibria dominate with respect to forecast improvement is strongly confirmed by the posterior probabilities for both period A and period B; the pooled probabilities (cols. 7–10) of 99.4 percent for PF and URE versus 1.7 percent and 14.7 percent for EPI and PI could hardly be more conclusive. The prediction that the strong-form equilibria provide more accurate forecasts overall, on the other hand, is definitely not supported; this prediction has an overall probability of only 37.1 percent (col. 13). Finally, the probabilities of 96.8 percent and 93.8 percent (cols. 14 and 15) lead us to accept the hypothesis that the strong-form equilibria provide superior price forecasts in the terminal year.

We now turn to an analysis of allocations. For each trading period and relevant forecast we can obtain the number of certificates "misallocated" and process these numbers just as we did the corresponding price RMSDs. Table 4 may be used to infer these misallocations, which are listed in table A2 in Appendix C. Again we predict that the strong-form forecast misallocations will decline over Market Years and that those for the weak-form forecasts will increase. We are also interested in the comparative accuracy of the alternative forecasts.

Table 6 presents the results of tests on our allocation predictions that are comparable to those just discussed for the price predictions. Pooling data over all experiments and both periods A and B, we attach 96.8 percent confidence to our prediction about strong-form forecast improvement (this refers to the PF forecast; the corresponding URE confidence level is 90.2 percent). Contrast this with the mere 5.7 percent confidence we can attach to EPI-forecast improvement (and 5.7 percent confidence to PI-forecast improvement). The only case in which the confidence of a weaker-form forecast improvement exceeds that attached to the stronger form is in period A of Experiment 5, but the confidence levels are generally low in this case for all of the forecasts. As for overall forecast accuracy, with the exception of Experiment 2 (in which the various allocation forecasts are observationally equivalent) and period B of Experiment 5, we find that the strong-form forecasts always dominate; the pooled probability of 99.7 percent in favor of this prediction is quite conclusive. We obtain equally strong results for terminal year accuracy, with pooled probabilities of strong-form superiority of 99.0 percent and 99.8 percent for periods A and B, respectively.

A final measure of convergence considered is the error in forecasting trading profits. Table 7 presents actual trading profits in each Market Year (defined as total payments to subjects less transactions commissions and dividends that would accrue if no trade took place) as a percentage deviation from *theoretical* trading profits arising in each equilibrium (again excluding implied commissions and profit from no trade).[23] The strong-form equilibria once more dominate. Average forecast error is only 10 percent for PF versus 259 percent for PI in Experiment 1 and 21 percent versus 15 percent in Experiment 2; it is 23 percent for URE versus 112 percent for EPI in Experiment 3, and 27 percent versus 46 percent in Experiment 4. Similarly in Experiments 5 and 6 we find that, conditional on leakage occurring or not occurring, the stronger-form forecasts win (compare PF and PI for "leakage" and URE and EPI otherwise). Pooling over all experiments, we have computed the confidence of profit forecast improvement as 19.3 percent for EPI, 10.6 percent for PI, 32.8 percent for URE, and 67.2 percent for PF. The confidence in the hypothesis that the stronger-form forecasts are more accurate over all years exceeds 99.9 percent.

The informational role of futures markets.—We now evaluate our second major prediction, that futures markets stabilize spot prices and enhance efficiency.[24]

A direct measure of price volatility is the coefficient of var
(CV) of transacted prices.[25] Table 8 lists the probability of a low
for spot prices and, separately, the probability of a greater (

---

[23] We measure theoretical profits by evaluating the payouts to traders if they
theoretically predicted allocations *and* received the theoretically predicted div
Thus each of these profits is simply the predicted period A equilibrium prices
aggregate number of certificates (18 in our experiments) Hence we are
measuring allocational efficiency, with misallocations weighted according to how much
they reduce gains from trade

[24] Harrison (1984) studies in detail the behavior of the three experiments with futures markets, drawing five conclusions (i) the presence of event uncertainty tends to retard the efficiency of asset markets, as measured by the extent to which prices and allocation reflect stronger-form equilibria, (ii) the presence of conclusive inside information encourages efficiency in an event uncertainty environment, (iii) the presence of event uncertainty and inside information has roughly offsetting effects on efficiency, although the effect of event uncertainty dominates slightly, (iv) event uncertainty is associated with a marked tendency to depart from routine hedging behavior; and (v) the presence of conclusive inside information encourages the adoption of routine hedges, even when event uncertainty is a joint treatment variable

[25] The CV is just the ratio of the standard deviation to the mean The objective of this transformation is to normalize the standard deviation in order to correct for parameter differences across experiments The CV measures are presented in table A3 in Appendix C and were derived from the data listed in figs. 1-6.

TABLE 7

PERCENTAGE DEVIATIONS OF ACTUAL PROFITS FROM PREDICTED EQUILIBRIUM PROFIT

EXPERIMENT AND YEAR	STATE	EQUILIBRIUM CONCEPT			
		EPI	PI	URE	PF
1					
1	...	...	158	...	−35
2	...	...	285	...	−4
3	...	...	250	...	−12
4	...	...	300	...	0
5	...	...	300	...	0
All		...	259	...	−10
2					
1		...	17	...	−20
2	...	...	28	...	−12
3	...	...	28	...	−12
4	...	...	−8	...	−36
5	...	...	9	...	−25
All	...		15		−21
3					
1	X	61	...	−35	...
2	Y	227	...	−7	...
3	X	95	...	−21	...
4	Y	241	...	−3	...
5	X	38	...	−44	...
X years	...	65	...	−33	...
Y years	...	234	...	−5	
All	...	112	...	−23	...
4					
1	X	10	...	−42	...
2	Y	113		−4	
3	Y	123	...	0	...
4	X	−33	...	−65	...
5	X	54	...	−19	...
X years	...	10	...	−42	...
Y years	...	118		−2	...
All	...	46	...	−27	...
5					
1	Y	422	−5	23	−20
2	Y	483	6	37	−10
3	X	40	−2	−41	−24
4	X	−5	−33	−60	−49
5	X	−11	−37	−63	−52
X years		452	1	30	−15
Y years	...	8	−24	−55	−42
All	...	107	−11	−26	−28
6					
1	X	−72	−55	−81	−73
2	Y	241	33	71	−15
3	Y	275	46	87	−6
4	X	−29	12	−53	−34
5	Y	300	56	100	0
X years	...	−50	−22	−67	−54
Y years	...	272	45	86	−7
All	...	90	29	10	−19

TABLE 8

EFFECT OF FUTURES MARKETS ON SPOT-PRICE VARIABILITY

Measure and Experimental Design	Experiments Compared	Period A	Period B	Period A and Period B
CV:				
Market uncertainty	1 and 2	.5	5	602
Event uncertainty	3 and 4	.845	845	947
Insider uncertainty	5 and 6	274	.206	.047
Pooled (overall)	(1, 3, 5) and (2, 4, 6)	.653	.702	720
Pooled (final years)	(1, 3, 5) and (2, 4, 6)	.889	421	803
CV improvement:				
Market uncertainty	1 and 2	5	0	333
Event uncertainty	3 and 4	.667	667	757
Insider uncertainty	5 and 6	833	667	971
Pooled	(1, 3, 5) and (2, 4, 6)	845	925	.944

NOTE.—Measures of each CV and CV improvement are listed in tables A3 and A4, respectively; the probabilities shown are based on a Mann-Whitney test.

provement in the experiments with a futures market than in those without one.[26] With pooled data we accept the hypothesis of a lower CV with futures markets at the 65.3 percent confidence level in period A and a 70.2 percent confidence level in period B. If we pool the final year data across all experiments and periods, we obtain an 80.3 percent confidence level for this hypothesis. Note, however, the sharp contrast in the results between the event uncertainty and insider uncertainty designs. Our results support with even more confidence the hypothesis that the presence of futures markets increases the improvement in CV (94.4 percent confidence overall). This result obtains for both event uncertainty and insider uncertainty designs.

Given the strength of our conclusion that the presence of futures markets reduces spot-price variability, we should clarify the relationship between our results and the results reported in FPP*b*. It can be shown (see table A5 in App. C) that futures markets in the FPP*b* study led to a significantly lower CV in spot prices in the terminal trading period (79.4 percent confidence) and to increased CV improvement (72.2 percent confidence). On the other hand, the evidence is (overwhelmingly) contrary to the hypothesis that FPP*b* futures markets reduced the CV over all trading periods (only 0.3 percent confidence). Note, however, that they employ the joint treatment variable of "futures markets *and* trader inexperience"[27] and do not isolate "futures markets" or "inexperience" as single treatments. The results of FHS, in which these two treatments were separately identified, suggest that each has qualitatively different effects on spot-price variability (with a higher CV being associated with "inexperience" and a lower CV with "futures markets"; see table 7 of FHS). We therefore conclude that, allowing for the minor design differences that may affect CV in the transition to equilibrium, the FPP*b* data also support our result that spot prices are less volatile in the presence of futures markets.

---

[26] A (one-tail) nonparametric Mann-Whitney test is used, as in cols. 11–15 of table 6. The probability measures show the confidence attached to the hypothesis that the presence of futures markets lowers the CV of spot prices or increases the proportional CV improvement. The latter probabilities are based on data shown in table A4 of Appendix C.

A summary measure of market trading efficiency is the profit earned relative to the stronger-form profit prediction (excluding, as before, commissions and no-trade profits). Table 9 shows the confidence we attach to the hypotheses that efficiency is greater in the presence of futures markets and, separately, that the improvement in efficiency is also greater with futures markets. Efficiency appears to be enhanced with futures markets in the insider uncertainty design (72.6 percent confidence), while there is no detectable effect under event uncertainty and indeed a reduction in efficiency under market uncertainty (see FHS for a detailed analysis of the comparative behavior of Experiments 1 and 2). The pooled results are similarly mixed, with the overall probability of 41.0 percent slightly unfavorable and the final-years probability of 65.5 percent slightly favorable to our prediction. The results are much more conclusive with respect to the effect of futures markets on the improvement in trading efficiency. Our prediction is accepted with 84.5 percent confidence overall and 83.3 percent confidence in each of the event uncertainty and insider uncertainty designs.

A more subtle effect of futures markets appears in connection with the comparison of strong and semistrong forms of efficiency in Experiments 5 and 6. From table 6, one can see that the semistrong price forecast (URE) did less well in the presence of a futures market (Experiment 6) than in its absence (Experiment 5) with respect to forecast improvement, while the strong-form price forecast (PF) performed equally well in the two experiments with respect to improvement and notably better in Experiment 6 with respect to overall accuracy. Similar comparisons for allocation forecasts also suggest that futures markets enhance the performance of strong-form forecasts relative to that of semi-strong-form forecasts.

A final piece of evidence is relevant in this connection. In table 10,

---

[27] The experiments reported in FPP*b* employed a strong intertemporal inseparability in preferences (i.e., the marginal value of a certificate held in period B depended strongly on how many certificates had been held in period A, and vice versa). It is not clear to us whether this design feature has an important effect on spot price variability.

TABLE 9

EFFECT OF FUTURES MARKETS ON TRADING EFFICIENCY

Measure and Experimental Design	Experiments Compared	Probability of Greater Efficiency
Trading efficiency.		
Market uncertainty	1 and 2	048
Event uncertainty	3 and 4	5
Insider uncertainty	5 and 6	726
Pooled (overall)	(1, 3, 5) and (2, 4, 6)	.410
Pooled (final years)	(1, 3, 5) and (2, 4, 6)	.655
Improvement in trading efficiency.		
Market uncertainty	1 and 2	.5
Event uncertainty	3 and 4	833
Insider uncertainty	5 and 6	833
Pooled	(1, 3, 5) and (2, 4, 6)	845

NOTE.—Measures of trading efficiency and improvement in trading efficiency are listed in tables 7 and A6 respectively, the probabilities shown are based on a Mann Whitney test

we compare the average trading profits of insiders to noninsiders in Experiments 5 and 6.[28] One can see that the average insider did indeed earn higher profits than the average outsider in every Market Year. More interestingly, however, insiders did notably less well when a futures market was present: a Mann-Whitney test allows us to associate 72.6 percent confidence with this claim.

V. Conclusions

Our experimental results support the following broad conclu

1. Market outcomes evolve over time toward strong-form in tionally efficient equilibria. This finding is quite robust to th ence or absence of futures markets and/or event uncertainty evident in our analysis of prices, allocations, and profits.

2. The presence of futures markets tends to stabilize spot pric measured by the coefficient of variation, transacted spot prices were typically less volatile in experiments that employed futures markets than in companion experiments with no futures markets. The effect

---

[28] Profits were normalized as usual by deducting security (no-trade) profit and ignoring transactions commissions For outsiders, security profits in period A are measured by taking the dividends prevailing in the worst possible state of nature and multiplying them by initial certificate endowments Security profits in periods B and C are calculated using the actual dividends, since the state of nature is revealed (this is the measure of security, or no-trade, profits employed in table 7) Security profits for insiders could be measured in exactly the same manner, and we show the results of doing this in the first column. An alternative measure that reflects their inside knowledge evaluates endowments in period A (and B and C) using actual dividends.

TABLE 10

INSIDER PROFITS AS A RATIO OF AVERAGE OUTSIDER PROFITS

EXPERIMENT AND YEAR	STATE	INSIDER SECURITY PROFIT MEASURE: Same as Outsiders	INSIDER SECURITY PROFIT MEASURE: Different from Outsiders
5			
1	Y	3.89	4.26
2	Y	1.57	1.71
3	X	3.49	3.86
4	X	1.18	1.30
5	X	1.62	1.78
6			
1	X	1.23	1.32
2	Y	1.34	1.69
3	Y	3.16	4.03
4	X	1.37	1.47
5	Y	2.80	3.54

was particularly strong in the event uncertainty design and was significant overall.

3. The presence of futures markets tends to speed the evolution of markets to more informationally efficient equilibria, at least in the presence of event uncertainty. The greater complexity and shorter trading time confronting our traders in experiments with futures markets seemed to impair market efficiency at first, but efficiencies in the terminal year were comparable to those in the corresponding experiments with no futures markets.

4. Futures markets promote the "leakage" of insider information. When a futures market was present we found that insiders earned significantly lower profits and that "leakage" equilibrium prices were better forecasts of actual transacted prices. In particular, strong-form informationally efficient forecasts clearly outperformed semi-strong-form forecasts *only* when a futures market operated.

These conclusions must, of course, be confirmed by other experimenters before they can be regarded as fully established. Moreover, several of our specific conclusions rest on a small sample (several Market Years or one experiment, depending on one's definition of the "sample"). In some ways our experimental design is the most complex so far employed, and it is comforting that our first conclusion corresponds so well to the findings of most other experimenters. On the other hand, we have noted that our second and third conclusions differ in some respects from the findings of FPP*b* and others, while our last conclusion seems entirely new. New experiments are called for, and it seems particularly important to find designs that minimize the possibility of trader confusion and information overload. Computer-mediated markets should be very helpful in this respect.

Appendix A

A Theoretical Model

In this Appendix we provide an informal game-theoretic justification for the various equilibria computed in Section III*B* of the text. We briefly summarize the relevant aspects of the theoretical model in Friedman (1982) and sketch the extension to multiperiod asset markets and event uncertainty.

Agents, often referred to as *traders*, are indexed by subscripts $i = 1, \ldots, 9$. They deal in *assets* indexed by superscripts $k = 0, 1, 2, 3$. Asset 0 is called *cash* or *money* and is divisible. The other assets represent dated certificate holdings, or claims on dividends—that is, asset 1 is a claim on period A dividends, acquired by holding a certificate at the end of the period A round of trading; similarly for assets 2 and 3, with respect to period B and C dividends. Note that assets 1–3 are not divisible—they can only be held in nonnegative integral quantities. The market exists over some finite time interval $[0, T]$ representing a Market Year.

We characterize each agent i by his

a) *endowment* or *initial holdings*, a nonnegative vector $\mathbf{x}_i(0) = x_i^0(0), x_i^1(0), x_i^2(0), x_i^3(0)$, the last three components of which are integral;

b) *preferences* represented by a smooth and monotone utility function $U_i[x_i(T), z]$, defined over terminal holdings $\mathbf{x}_i(T) = x_i^0(T), \ldots, x_i^3(T)$ and the state of the world $z \in Z$; and

c) *information*, consisting of an initial private information set $I_i(0)$ together with public information available at time t, $I(t)$.

We partition the Market Year $[0, T]$ into subintervals $[0, T_A]$, $(T_A$ $T_C]$ called *trading periods*, where $0 < T_A < T_B \leq T_C = T$, and specify bundles y^l of assets 1–3 that can be exchanged for money in each period. For instance, in the spot-only markets, we have $T_B < T_C$ (th riods); the only bundle traded in the first period is a package y^A of sh of assets 1, 2, and 3; in the second period it is a similar package of y^B 2 and 3, and in the third it is simply asset 3, $y^C = x^3$. In the futures experiments, $T_B = T_C = T$ (two periods), and single units of asset 3 can traded in both periods in addition to the packages mentioned previously

The form of trading is *double auction:* at each time $t \in [0, T]$, each agent i may announce an ask price $p_i^{al}(t)$, or bid price $p_i^{bl}(t)$, representing the amount of money he is prepared to accept, or give up, in exchange for a unit of a package of y^l that can be exchanged at time t. If another agent j chooses to accept i's offer (or bid) at time $s \geq t$, then holdings of i and j change accordingly: $\Delta x_i(s) = -\Delta x_j(s)$, where $\Delta x_i^0(s) = p_i^{al}(t)$, and $\Delta x_i^k(s) = -1$ for k in the package l in the case of an offer, and $\Delta x_i^0(s) = -p_i^{bl}(t)$ and $\Delta x_i^k(s) = 1$ in the case of a bid. For all other times s', assets k', and agents i', $\Delta x_{i'}^{k'}(s') = 0$. Once made, an agent's offer (or bid) may be accepted by any other agent and must be honored until it is superseded by a new offer (new bid) by the same agent or a lower offer (higher bid) by another agent, or the trading period ends. The last accepted price at or before time t is denoted $p^l(t)$. We refer to the lowest offer (highest bid) as the *market* offer (bid).

Our basic behavioral assumption is that agents attempt at each time t to maximize expected utility $E_{it}U_i[x_i(T)]$, where E_{it} is the conditional expectation operator employing probabilities (regarding states of the world and trading opportunities) based on i's current information set $I_i(t) = I_i(0) \cup I(t)$. The public information set $I(t)$ consists of all market bids and offers and all acceptances made at times $s < t$. Various assumptions regarding the private information sets $I_i(0)$ will be given later.

A strategy S_i for agent i is a decision rule that assigns to each possible information set $I_i(t)$ and each tradable bundle l bid and offer prices $p_i^{al}(t)$ and $p_i^{bl}(t)$ and acceptance prices $v_i^{al}(t)$ and $v_i^{bl}(t)$. We make the convention that i announces his bid (offer) at time t only if it betters the current market bid (offer) or if he currently already holds the market bid (offer). As for the v's, we assume that i will accept any other agent's bid if and only if it equals or exceeds $v_i^{al}(t)$ and any offer if and only if it is no smaller than $v_i^{bl}(t)$, with $p_i^{al}(t) \geqslant v_i^{al}(t) \geqslant v_i^{bl}(t) \geqslant p_i^{bl}(t)$. (Friedman [1982] shows that these assumptions regarding acceptance prices involve very little loss of generality.)

Finally, a strategy S_i is admissible if it ensures that the nonnegativity constraints $x_i(t) \geqslant 0$ will always be satisfied. The following restrictions suffice to ensure admissibility: if $x_i^k(t) = 0$ for some asset k in the package l, then $v_i^{al}(t) = B$, where B is a very large number (say, the total endowment of money), and if $x_i^0(t) = 0$, then $v_i^{bl}(t) = 0$, for all l.

Before discussing and computing equilibria of our model, we note some special features of our experimental parameterizations that will simplify the analysis. First, for each agent i and each state of the world z, the marginal rates of substitution of assets 1–3 for money are *constant* and equal to the corresponding "dividend yields." Hence, each agent's utility depends on terminal asset holdings $x_i(T)$ only through its cash equivalent $w_i(T, z) = d_i(z) \cdot x_i(T) = 1x_i^0(T) + d_i^1(z)x_i^1(T) + d_i^2(z)x_i^2(T) + d_i^3(z)x_i^3(T)$, where $d_i^k(z)$ is i's per certificate dividend at the end of the kth trading round in state z for $k = 1, 2, 3$. Thus we may abuse notation slightly and write $U_i[x_i(T), z] = U_i[w_i(T, z)]$—that is, we regard i's preferences as defined over terminal wealth and can employ standard concepts of risk aversion, etc.

The second special feature is that none of our agents has unique endowment and dividend parameters—each has two "clones." This will reduce the possibility of monopolistic behavior, as we shall see. Of course, even our clones may differ in ways other than the experimentally induced parameters such as in risk aversion or learning speed.

A final feature of our experiments is that agents are all endowed with sufficient cash that the nonnegativity constraint on x_i^0 may safely be ignored.

Equilibria of the Model

Several different notions of equilibrium are relevant to our model. First, there is Walrasian equilibrium: abstracting from strategic considerations, real-time trading, incomplete markets, and incomplete information, we may compute prices $p^*(z)$ and final asset allocations $x^*(z)$ that would arise if the state of the world were known to be z and agents maximized utility (subject to indivisibility and nonnegativity constraints) on budget sets defined relative to market-clearing prices. Similar computations, performed near the end of this subsection, can also yield Walrasian prices $p^*(\Pi)$ and allocations $x^*(\Pi)$ that would result from an expected utility maximization if all agents are risk neutral and share common probabilistic beliefs Π over states of the world. For reasons that will be apparent by the end of this subsection, we refer to $[p^*(z), x^*(z)]$ as a perfect-foresight (PF) equilibrium and $[p^*(\Pi), x^*(\Pi)]$ as an uninformed rational expectations (URE) equilibrium.

Double-auction markets such as ours, however, inevitably involve strategic considerations. Each agent must pick a real-time bid, offer, and acceptance strategy S_i and should realize that the outcome of following S_i depends on others' strategies S_j. Hence in general an optimal strategy for agent i is an admissible strategy that maximizes expected utility of final asset holdings, relative to possibly incomplete information regarding others' strategies and states of the world. We shall impose the consistency condition that, by the end of each trading period, traders' actions (choice of p's and v's) are in (ex ante) Nash equilibrium. That is, each trader's final action is an "expected best response" against other traders' final actions, given nondogmatic (but not necessarily mutually consistent) beliefs about others' intentions to accept at $t = T_A$ (respectively, T_B, T_C). The idea is that traders do *not* jam the market by all waiting until period end to seize the market bid or ask or to accept market prices; see Friedman (1982, 1984) for further discussion and Glover (1983) for some theoretical justification of this no-congestion assumption. A selection of optimal strategies consistent with no congestion is referred to as a temporary noncooperative equilibrium (TNE). It is intuitively clear, however, that the hypothesis of TNE places very few restrictions on the course of trade in our model and therefore is not very useful unless we make quite specific assumptions regarding agents' information.

Two polar sets of assumptions seem particularly interesting. In the first, we assume that agents initially know only their own parameters, have quite diffuse priors regarding others' strategies, and are quite risk averse, so they follow maximin strategies. We refer to this as the private information (PI) case.

PROPOSITION 1: In the case of no event uncertainty and spot markets only (Experiment 1), the private information assumptions imply that the closing market bid and ask prices for certificates are $p(\mathrm{A}|\mathrm{PI}) = \max_i \Sigma^3_{k=1} d^k_i$ in period A, $p(\mathrm{B}|\mathrm{PI}) = \max_i \Sigma^3_{k=2} d^k_i$ in period B, and $p(\mathrm{C}|\mathrm{PI}) = \max_i d^3_i$ in period C

A formal proof can be obtained from minor adaptations of proposition [illegible] in Friedman (1982), so we will just sketch the argument here. For t [illegible] agent i increases his wealth $w_i(t)$ if he sells a certificate at a price exc[illegible] "security level" $v^*_{iA} = \Sigma^3_{k=1} d^k_i$ or buys at a price below v^*_{iA} The PI ass[illegible] imply that agents will not hold out for better prices in later periods [illegible] double-auction trading rules require that an agent accept the mar[illegible] (bid) or else better the market bid (offer) if he wishes to buy (sell [illegible] (dropping the index l since only the bundle $l = \mathrm{A}$ can be sold) we h[illegible] $v^b_i(t) \rightarrow v^*_{iA}$ as $t \rightarrow T_A$ for all i,[29] so the closing market bid is $\max_i v^*_{iA}$ [illegible] Likewise, $p^a_i(t)$, $v^a_i(t) \rightarrow v^*_{iA}$ if $y_i(t) > 0$ (i.e., if i can still sell) as $t \rightarrow T_A$, but i [illegible] sell all his certificates so $y_i(T_A) = 0$ if v^*_{iA} is not maximal Hence the market bid coincides with the market ask. The argument for period B is exactly similar, with $v^*_{iB} = d^2_i + d^3_i$ as the eventual reservation prices. The period C argument is a direct consequence of Friedman (1982, proposition 4.1) for the special case of no liquidity constraints and "clones"; the PI assumptions are not required here.

COROLLARY: Under the assumptions of proposition 1, the closing *allocations* satisfy $y_i(T_\alpha) = 0$ if $v^*_{i\alpha} < p(\alpha|\mathrm{PI})$, $\alpha = \mathrm{A, B, C}$.

Proposition 1 suggests the prices and allocations that would arise in the first Market Year if all traders are cautious but sophisticated; this is the situation that the PI assumptions are intended to capture. Note that transacted prices below $p(\cdot \mid \mathrm{PI})$ early in the trading round are consistent with the PI assumptions but higher transacted prices are not: buyers who pay higher prices have better information than assumed, are willing to gamble, or have made a mistake.

---

[29] At least for all i who have sufficient cash, but we have assumed that everyone does. The assertion also depends heavily on the "clone" assumption because otherwise an agent can hold the market bid below v^*_{iA}

The existence of a futures market provides additional information to traders, so proposition 1 must be modified for this case. As was argued informally in Section III*A* of the paper, traders can "lock in" a resale price for period C, and by the last part of proposition 1 above, this price should converge to $p(\mathrm{C} \mid \mathrm{PI}) = \max_i d_i^3$. Hence for Experiments 2 and 4, $p(\mathrm{A} \mid \mathrm{PI}) = p(\mathrm{C} \mid \mathrm{PI}) + \max_i (d_i^1 + d_i^2)$ while $p(\mathrm{B} \mid \mathrm{PI}) = p(\mathrm{C} \mid \mathrm{PI}) + \max_i d_i^2$, and the PI allocations are modified accordingly.

Given the lack of information under *PI* regarding resale possibilities in subsequent trading rounds, it seems reasonable to assert that $v_i^b(t) \rightarrow v_{iA}^* = \Sigma_{j=1}^3 d_i^j$ as $t \rightarrow T_A$ for our spot-only experiments with no event uncertainty. However, maximin strategies under event uncertainty would yield $v_i^b(t) \rightarrow \min_{z \in Z} \{\Sigma_{j=1}^3 d_i^j(z)\}$, which seems altogether *too* cautious given *i*'s well-defined priors over states of the world *z* in our Experiments 3–6. Hence, in analyzing Experiments 3 and 4, we examine expected private information (EPI) strategies, in which traders share firm priors Π over states of the world and undertake only trades that raise current expected wealth. One can think of such traders as extremely risk averse with respect to Knightian uncertainty and risk neutral with respect to Knightian risk. Perhaps a better interpretation is that, under event uncertainty, $v_{iA}^*(\Pi) = \Sigma_{j=1}^3 \Sigma_{z \in Z} \Pi(z) d_i^j(z)$ is the appropriate notion of *i*'s "security level" in period A trading. In any case, we obtain by arguments analogous to those of proposition 1 that $p(\mathrm{A} \mid \mathrm{EPI}) = \max_i v_{iA}^*(\Pi)$ is the closing market bid and ask price in period A, and $x_i(T_A) = 0$ if $v_{iA}^*(\Pi) < p(\mathrm{A} \mid \mathrm{EPI})$, for the spot-market-only Experiment 3. For Experiment 4, the existence of a futures market raises the EPI price to $p(\mathrm{A} \mid \mathrm{EPI}) = p(\mathrm{C} \mid \mathrm{EPI}) + \max_i \Sigma_{j=1}^2 \Sigma_{z \in Z} \Pi(z) d_i^j(z)$, where $p(\mathrm{C} \mid \mathrm{EPI}) = \max_i \Sigma_{z \in Z} \Pi(z) d_i^3(z)$ is the closing price of futures contracts in period A. EPI allocations of certificates and futures contracts satisfy the usual condition. For period B and C trading (both spot and futures), there is no event uncertainty in our experiments because the state of the world *z* is revealed at $t = T_A$. Therefore, the previously computed PI prices and allocations apply to those periods with the convention that $d_i^j = d_i^j(z)$. Such prices are denoted $p(\cdot \mid \mathrm{PI}, z)$.

Finally, for Experiments 5 and 6 where the state of the world *z* is known at $t = 0$ by some traders, we note that $p(\mathrm{A} \mid \mathrm{EPI})$ are the relevant benchmark prices at $t = T_A$ if insiders are able to keep their information from leaking out (since insiders do not have clones), while $p(\mathrm{A} \mid \mathrm{PI}, z)$ apply if the information does leak.

Let us now turn to the polar case in which strategies are based on excellent prior information; in fact, we will assume that agents know *everything* that is relevant. We define a *rational noncooperative equilibrium* (RNE) as a TNE strategy selection $S = (S_1, \ldots, S_9)$, which is stable in the sense that *i* cannot expect to gain by departing from S_i if he expects others to follow S_j, and he cannot actually gain by departing from S_i if others in fact follow S_j. Formally, *S* is an RNE if, for all *i*, $E_{i0}U_i[w_i(T) \mid S \backslash S_i']$ is maximized at $S_i' = S_i$ and $U_i[w_i(T) \mid S_i']$ is maximized at $S_i' = S_i$ if S_j is employed for all $j \neq i$.[30]

Proposition 2: In the case of no event uncertainty and spot markets only (Experiment 1), there is an RNE strategy selection *S* such that *all* transactions take place at perfect-foresight (i.e., Walrasian) prices $p(\cdot \mid \mathrm{PF})$, that is,

---

[30] $S \backslash S_i'$ means $(S_1, \ldots, S_{i-1}, S_i', S_{i+1}, \ldots, S_9)$. Maximization is over *i*'s set of admissible strategies.

$$p(t) = p(\mathrm{A} \mid \mathrm{PF}) \equiv \sum_{k=1}^{3} \max_i d_i^k = \sum_{k=1}^{3} p^{*k}, \quad \text{for } t \in (0, T_\mathrm{A}],$$

$$p(t) = p(\mathrm{B} \mid \mathrm{PF}) \equiv \sum_{k=2}^{3} \max_i d_i^k = \sum_{k=2}^{3} p^{*k}, \quad \text{for } t \in (T_\mathrm{A}, T_\mathrm{B}],$$

$$p(t) = p(\mathrm{C} \mid \mathrm{PF}) \equiv \max_i d_i^3 = p^{*3}, \quad \text{for } t \in (T_\mathrm{B}, T_\mathrm{C}].$$

Again, this proposition may be proved formally by minor adaptations of a proof in Friedman (1982, proposition 4.2), so we just sketch the argument here. First, one verifies that $p^* = (p^{*1}, p^{*2}, p^{*3})$ are the unique Walrasian prices for $p^{*k} = \max_i d_i^k$, $k = 1, 2, 3$, and that the corresponding allocations of assets satisfy $x_i^{k*} = 0$ if $d_i^k < p^{*k}$—this verification is quite straightforward given that liquidity constraints are not binding. Then one notes that if each agent follows the strategy of buying (selling) certificates only at prices at or below (above) their remaining value relative to p^*, but always buys (sells) at such prices if it increases his wealth,[31] then the perfect-foresight price system will prevail as asserted in the proposition. Finally, one verifies routinely that such a strategy selection is an RNE.

COROLLARY: Under the RNE strategy of proposition 2, the allocations satisfy $y_i(T_\mathrm{A}) = 0$ if $d_i^1 < p^{*1}$, $y_i(T_\mathrm{B}) = 0$ if $d_i^2 < p^{*2}$, and $y_i(T_\mathrm{C}) = 0$ if $d_i^3 < p^{*3}$, that is, certificates are all held at the end of each trading period by agents with maximal dividends for that period.

Note that $p(\cdot \mid \mathrm{PF}) \geq p(\cdot \mid \mathrm{PI})$ and is usually strictly greater. For instance, for the parameters of table 2, $p(\cdot \mid \mathrm{PF}) = (1.85, 1.15, 0.70)$ and $p(\cdot \mid \mathrm{PI}) = (1.35, 1.00, 0.70)$.

Analogous prices and allocations can be derived for our other experiments. The existence of futures markets is irrelevant if agents are fully [illegible] already, so the same PF prices and allocations apply to Experim[illegible] Experiments 3 and 4 we use the notion of uninformed rational e[illegible] (URE): all agents are risk neutral and know the probabilities [illegible] possible states of the world and the RNE prices $p(\mathrm{C} \mid \mathrm{PF}, z) = \max$ [illegible] $p(\mathrm{B} \mid \mathrm{PF}, z) = \max_i d_i^2(z) + p(\mathrm{C} \mid \mathrm{PF}, z)$ that will obtain in subseque[illegible] once the state z is revealed. Then we can replace $p(\mathrm{A} \mid \mathrm{PF})$ by $p($A [illegible] $\max_i \Sigma_{z \in Z} \Pi(z)[d_i^1(z) + p(\mathrm{B} \mid \mathrm{PF}, z)]$ as the full-information benchma[illegible] experiments, because under these circumstances agents will be willing [illegible] up to their expected dividend in period A plus expected resale value [illegible] period B. For Experiment 4 the relevant period A futures price is $p(\mathrm{C} \mid \mathrm{URE})$ $= \Sigma_{z \in Z} \Pi(z) p(\mathrm{C} \mid \mathrm{PF}, z)$, while the period B futures price reverts to $p(\mathrm{C} \mid \mathrm{PF}, z)$. Finally, for Experiments 5 and 6, we note that the URE prices (and corresponding allocations) apply if information does not leak from insiders,[32] but PF prices apply if there are leaks.

The efficient PF (or URE) outcomes should be thought of as arising after traders have learned everything relevant, after many Market Years. To justify this interpretation, we must argue that traders can in fact learn what they need to know. In markets run by a Walrasian auctioneer, this may not be possible. But in repetitively stationary double-auction markets, agents acquire considerable information during the trading process; Friedman and Harrison (1984) show how a simple Bayesian learning scheme suffices for the market to converge to a Walrasian equilibrium outcome.

---

[31] If he has any certificates to sell.

[32] Again, the fact that a single clone of each type has inside information is crucial here.

Appendix B

Market Commentary

This commentary is designed to complement the descriptive account of our experimental results provided in figures 1–6 (price behavior) and table 4 (allocation behavior).

It is convenient to refer to periods A, B, and C by the notation PA, PB, and PC, respectively. Similarly, we refer to years 1, 2, 3, 4, and 5 as Y1, Y2, Y3, Y4, and Y5, respectively. All other abbreviations are as defined previously (e.g., PI, PF, EPI, URE). Copies of the raw experimental data on individual transactions are available on request from the second author.

Experiment 1

Period C

Closing prices in Y2 indicate to the traders that someone is prepared to pay nearly $0.70 for a certificate in PC. This signal is confirmed in Y3, Y4, and Y5 with increasing precision

Allocations converged to the equilibrium prediction quickly, with 14/18 (read as "14 certificates out of 18 predicted to be held by the correct trader type"), 18/18, 16/18, 18/18, and 18/18 being held in each successive year. Note that the PF and PI predictions (for prices and allocations) are identical in PC.

Period B

With the closing prices in PC of Y2 near to their equilibrium value, PB prices in Y3 begin to move above their PI equilibrium value. The $0.05 increase in average PC prices from Y2 to Y3 (with a slight fall in variance also) provides the basis for PB prices in Y4 to converge dramatically to their PF equilibrium value. In Y5 we observe complete convergence in PB prices, following the near-complete convergence of PC prices in Y4. These results support the "swingback hypothesis" advanced by Grossman (1978) and studied experimentally by FPPa and FHS.

Allocations converge steadily to the PF prediction 3/18, 5/18, 16/18, 18/18, and 18/18 in Y1–Y5, respectively. The alternative PI prediction accounts for 14/15, 11/13, and 0/2 of the misallocation in Y1, Y2, and Y3.

Period A

Prices in PA begin in Y1 around the PI prediction but with a rather large dispersion. Note that PA prices in Y2 and Y3 are significantly above the PI prediction, despite the fact that previously observed PC prices had not converged to their PI equilibrium value. This is explained quite simply: given that the traders were expecting the market to be stationary from year to year, trader types I and II had seen PC prices in Y1 and Y2 that were well above their own private valuations ($0.45 and $0.10 per certificate, respectively). Without knowing at this stage how much higher PC prices might go, they could nonetheless embody these observed PC market valuations in their reservations prices in PA and PB. Thus we observe that the "swingback" from PC prices to PA valuations is not postponed until PC prices have stabilized; rather, it occurs concurrently.

Allocations are exactly as predicted by the PF concept from Y2 on, with the entire misallocation of four units in Y1 consistent with the PI prediction. Note that the allocational convergence in PA was more rapid than in PB.

A more detailed account of the behavior of this market is presented in FHS (reported there as Experiment 3).

Experiment 2

Period C (Futures)

In Y1 futures prices are already above the private PC valuations of trader types I and II ($0.10 and $0.45, respectively). By the end of PB in Y2 they are near to their equilibrium value, and from PB in Y3 on they provide a persistently clear signal.

Although trades in futures generally follow spot trading in each period in Y3 and Y4, there were numerous unaccepted bids and offers on futures at the time that spot contracts were being made. These bids and offers often contain useful information, despite not resulting in trades.

It is useful to examine the eventual PC holdings of spot certificates (these are just the delivered futures contracts pending at the end of PB). In Y1, Y2, and Y3 there were no misallocations in these holdings. In Y4 and Y5 some 13/18 and 15/18 were correctly allocated relative to both PI and PF concepts, which had identical predictions in this case. The misallocation in Y4 is discussed later.

Period B (Spot)

Despite the futures prices observed in PA and during PB, in Y1 we se
PB spot prices are just below their PI equilibrium value. Significant l
apparently occurs in Y2, and PB prices are at their PF equilibrium val
and Y4.

In Y5 of this experiment we randomly reassigned agents to differe
dend profiles without altering the aggregate market parameters. Th
here is the ability of agents to distinguish market signals from their
signal (namely, their own dividend profile)—the essence of our com
equilibrium notions. Despite a "technical reaction" of sorts, especially in
PA spot market, the results essentially repeat the behavior in Y1. The fact that we observe convergence to the PF equilibrium price in Y4 indeed tells us that agents had made the distinction between these two signals when forming their trading strategies; our results in Y5 confirm this conclusion.

The allocations of PB spot certificates in each year were 13/18, 18/18, 18/18, 11/18, and 16/18 relative to the PF prediction. All but one of the misallocations in Y1 are explained by the alternative PI prediction.

The misallocation in Y4 is attributable entirely to one type I agent (agent 5). Her spot holdings at the end of PA in Y4 were seven certificates, and she had mechanically hedged by selling seven futures: impeccable PF behavior so far. (In Y3 her only PA transaction was the purchase of one spot certificate, and she had sold all three spot certificates in PB as implied by PF behavior.) From the sequence of unaccepted bids and offers it appears that she was determined to buy in her futures position in PB before selling off her spot holdings; recall that we did not allow net short positions at any stage of trading during any period and that she had fully hedged in PA. There was a marked pattern in the sequence of trading in this experiment once prices were near to their PF values. The period would open with one or two errant (spot and futures) price signals away from the established equilibrium value, these bids or offers would go unaccepted (and were sometimes greeted with outward signs of amusement by other traders), and trading in spot and then futures would settle down at the previously established prices. Virtually all traders chose to establish their spot position and *then* enter the futures market (this is definitely not the observed pattern in earlier Market Years in which the transition from PI to PF prices is incomplete; in these years there is a healthy mixture of concurrent spot and futures trading). Given this observed pattern in Y4, agent 5 was "locked in" to her spot position by an inability to buy in her futures position until the latter part of the period. She was able to buy five futures during PB but was not able to sell any spot. However, she made no serious effort to extricate herself from her self-imposed corner by offering to buy futures for more than $0.80 and then offering to sell spot for less than $1.25 (the established PF prices).

Period A (Spot)

Given the futures prices available, PA spot prices as early as Y1 are well above their PI equilibrium value. Convergence to the PF value is nearly complete by the close of Y3 trading and is complete by Y4. The "technical reaction" in Y5, mentioned above, involved one trader aggressively bidding for spot certificates in PA at $2.05, some $0.05 higher than market signals would justify (given risk neutrality). Assessing her situation at the close of PA trading, the market "allowed" her to lighten her spot holdings in PB at a discount (namely, some $0.05 below the PF price in PB). This is the same agent (agent 5), incidentally, who tried to "stonewall" the market in PB in Y4.

Allocations in each year were 13/18, 13/18, 13/18, 18/18, and 14/18 relative to the PF prediction. The alternative PI prediction accounts for 4/5, 5/5, 5/5, 0/0, and 4/4 of the implied deviations in each year.

A more detailed account of the behavior of this market is presented in FHS (reported there as Experiment 4).

Experiment 3

Experiment 3 was the first of our event uncertainty markets. The equilibrium prices in PB and PC were independent of the state of nature, but PA maximal dividends were contingent. The PC allocation predictions were contingent, as were the PB predictions for the PI equilibrium concept.

Period C

Prices converged to near the equilibrium value by the end of Y1, and from Y3 on there was a clear and strong signal as to likely PC market valuations. Allocations in each year were 14/18, 18/18, 18/18, 18/18, and 11/18 relative to the equilibrium prediction for the relevant state of nature. The misallocation in Y5 of seven certificates appears to be due to one agent's being caught in a "musical chairs" situation midway through a block-churning trading process.

Period B

Recall that there were fair to excellent PC price signals from the close of Y1 to Y5. Apart from the very first transaction, virtually all PB prices were between the PI and PF predictions. In Y2 the PB prices were much closer to the PI value than in Y1 or Y3; this is consistent with Y2 being the first change in the state of nature. In PC of Y2, however, traders were able to learn that PC prices appeared to be independent of the state, a useful piece of information when forming their PB reservation price trading strategies.

In Y3, the second X year, observed PB prices were much closer to the PF prediction, indicating that the market had learned from the first two years. Prices were very close to the PF prediction in Y5, the third X year For the X years, Y1, Y3, and Y5 allocations were 17/18, 14/18, and 18/18, respectively, relative to the PF prediction. The implied misallocations in Y1 and Y3 are consistent with the alternative PI prediction.

There was dramatic intraperiod market learning in Y4, the second Y realization. For the Y years Y2 and Y4, allocations were 5/18 and 15/18, respectively, relative to the PF prediction In Y4 the deviation was due to one type III agent who was tardy in divesting himself of seven units held at the end of PA; the market was firm at $0.69, so he could have sold easily at a few cents lower.

Period A

There were good PB signals only prior to PA in Y5. After the first trade in Y1, all PA prices were well above the EPI equilibrium value, despite the poor PB signals. There is a steady improvement toward the URE prediction in the X years (Y1, Y3, and Y5) and Y years (Y2 and Y4) We discuss these dynamic paths in Section IV*B*

In the three X years, allocations were 5/18, 10/18, and 18/18 relative to the URE prediction; in the two Y years they were 14/18 and 11/18 All of the implied misallocations are consistent with the alternative EPI prediction

Experiment 4

The parameters of Experiment 4 were qualitatively similar to those of ment 3 with respect to the contingency of equilibrium predictions rium prices in PB and PC were independent of the state of nature equilibrium prices were contingent The futures certificate allocation tions were contingent during PB trading, and PB spot allocation pre were also contingent for the PI equilibrium concept. The random reali of the state of nature provided a change of state in the first two years, was also the case in Experiment 3.

Period C (Futures)

With uncertainty about the true state of nature in PA, futures prices in Y1 were at the EPI prediction. In Y2 several bids were unaccepted at that value and finally were several trades closer to the URE prediction. By Y3, in PA futures prices were at the URE value. There were no futures trades in PA of Y5, but there was no informational requirement for any trades (i.e., the nonacceptance of several bids just below the equilibrium value was a sufficient signal that the PC market valuation had not changed). After the first three trades in PB in Y1, virtually all PB futures prices in each year were at the predicted equilibrium level.

Spot allocations resulting in PC from PB futures positions were 15/18, 11/18, and 17/18 for the three X years (Y1, Y4, and Y5) relative to the equilibrium prediction and 18/18 for each of the two Y years (Y2 and Y3). The misallocation in Y4 is discussed later.

Period B (Spot)

There was a steady but slow movement of PB spot prices from the PI to the PF equilibrium prediction. This sluggish convergence occurred despite clear futures prices in PB as early as Y1 and the absence of any event uncertainty in PB.

Allocations showed a slow improvement to the PF prediction; they were 13/18, 11/18, and 17/18 in the X years and 12/18 and 16/18 for the Y years. Virtually all of the misallocations in the Y years are explained by the alternative PI prediction. The severe misallocation in Y4 was due to one type I agent who failed to unload seven certificates, losing $0.60 on each one he failed to sell given prevailing prices.

Period A (Spot)

Prices were below the EPI value in Y1, at the EPI value in Y2, and very slowly moving to the URE value in later years. The PC futures price signal was quite precise from Y3 on, although the PB spot signals were not as clear.

There were severe misallocations relative to the URE prediction, which was only correct in 6/18, 11/18, 9/18, 10/18, and 8/18 in each successive year. The alternative EPI prediction accounts for 5/12 and 5/7 of the misallocations in Y1 and Y2, respectively, but for none of the misallocations in the last three years. An explanation of these deviations as the result of event uncertainty is possible.

One implication of the presence of event uncertainty is the need to form some estimate of PC and PB prices for each state before forming reservation prices for PA trading. In the absence of event uncertainty such estimates could be possible as early as Y2 (even with a PC futures market in Y1, traders need to see PB spot prices for PA valuation purposes). In the presence of event uncertainty such estimates are only possible after traders have observed an X *and* a Y year, which could have been much later in the experiment if at all (in fact, it had occurred before Y3 in Experiments 3, 4, and 6). The necessarily greater amount of initial learning required to form URE reservation prices in PA (at least two Market Years, compared to one Market Year) in the presence of event uncertainty opened up numerous profitable disequilibrium trading strategies for certain traders that would be nonprofitable at equilibrium prices. An excellent example of the availability of such disequilibrium strategies is provided in this experiment and accounts for the severe misallocations, relative to our equilibrium predictions, that were noted earlier.

The main culprit in the misallocations in Y3, Y4, and Y5 was a particular type III agent who went on a buying spree for PA spot (contrary to EPI and URE predictions). In Y1 and Y2 PB spot prices averaged $1.618 and $1.660, respectively; even if a trader had not realized that PB market valuations were state independent, the expected PB price in Y3 would have been 0.5($1.618) + 0.5($1.660) = $1.639. Given this estimate for PB prices and assuming risk neutrality on our type III agent's behalf, his reservation price in PA would have been composed of 0.5($0.5) + 0.5($0.3) = $0.40 expected PA dividend plus a sale to market in PB for $1.639, giving a total reservation price in PA of $2.039. In fact he purchased at around $1.97, sorrowfully learned that this was a Y year, and collected his dividend of $0.30; he was then greeted with a PB spot market prepared to pay $1.706 instead of the $1.639 he had bargained for! Thus, despite Y3 being a "bad year" for him, this trader was able to survive with a few cents profit.

The same strategy in Y4 and Y5, with some adjustment upward in his PB estimate and hence his risk-neutral PA reservation price, earned him handsome profits (given that they were "good years" for him in terms of PA dividends). Note that he was able to profit from such a trading strategy only because PA spot prices were so "low." Thus the misallocations in question were not at all due to this type III agent's irrationality or myopia, but to market inefficiency in PA pricing. It was, in effect, the type I agents that were to "blame," for not buying actively enough in PA (a simple explanation for their inactivity could, of course, be a severe aversion to risk; thus we qualify the word "blame").

Experiment 5

Experiment 5 introduces insiders into an event uncertainty environment. Parameters were chosen such that PA, PB, and PC equilibrium prices were contingent (in Experiments 3 and 4, PB and PC price predictions were not contingent, although the allocation predictions were).

Period C

In Y1 and Y2, both Y years, prices were precisely at the equilibrium value. In the first X year, Y3, there were some problems: two trades $0.10 above the (new) equilibrium value, and the remainder $0.05 below the equilibrium value. For the final two years prices were close to the equilibrium value. The allocations were generally consistent with the equilibrium predictions for each state of nature, with 18/18, 16/18, 17/18, 16/18, and 16/18 correctly predicted.

Period B

By the end of PB in Y1 there was convergence to the PI equilibrium prediction. In Y2 prices opened at the PI value and slowly increased to midway between the PI and PF predicted values.

Y3 is the first X year, after two Y years, and PB prices were well abov
PF equilibrium given an X realization but *below* the PB prices in Y1 a
This behavior suggests a "hysteresis effect" and the absence of any leak
insider information. PB certificate prices in Y4 displayed a tendency sim
those in Y3, although they rapidly declined to below the relevant PF
rium value. In Y5 prices opened at the PI equilibrium value for an X
which was even lower than closing prices in Y4. Note that the PC spot
signal was quite precise in Y4 and yet was not embodied in PB valuations in Y5.

Allocations in PB converged toward the PF predictions in the two X years (1/18 and 18/18) and the three Y years (6/18, 14/18, and 16/18). Virtually all of the implied misallocations are explained by the alternative PI prediction, especially in Y1.

Period A

Prices in Y1 started below the EPI value and then clustered around it. In Y2 they opened at the EPI value and then moved toward the URE value. Prices in Y3 opened at the (new) PF equilibrium value for that year and quickly climbed to the URE value. This apparent "reverse learning" is consistent with opening prices' being due to hysteresis after the PA prices of Y2 (averaging $1.42, and with a PF value in Y2 of $1.45), especially with the observed convergence to the URE value. In Y4, prices were around the URE value, consistent with the absence of leakage of insider information. The bubble of insider secrecy obviously burst in Y5: prices opened at the PI value for an X year and quickly settled midway between the PI and PF equilibria, *both* of which were considerably below the URE value.

Experiment 6

The parameter values of Experiment 6 were qualitatively similar to those of Experiment 5 in terms of the contingency of theoretical predictions.

Period C (Futures)

Trading in PA of Y1 was at the EPI value. Surprisingly, however, it was the type III insider who bought in PA; his type is predicted to buy by the EPI rationale, but the insider knew that this was an X year. Perhaps realizing his error, he sold futures aggressively early in PB at low prices. The futures price eventually converged to the equilibrium value in PB.

A change in state occurred in Y2. In PA futures prices opened at the URE value, with the URE-predicted type III buying (only 2/6 to the insider) and type II selling (3/6 from the insider). Futures prices then converged to the equilibrium PF and PI value for a Y year, with traders possibly receiving a signal as to the true state of nature from PA spot prices (discussed further below). Type III bought (6/12 by the insider), as predicted by the PF *and* PI equilibria. Another Y realization occurred in Y3, and PA futures prices opened at the URE value. The type III outsider was buying from a type II insider at these prices. Consistent with leakage of insider information with these trades, futures prices in PA rapidly converged on their PF and PI equilibrium value for a Y year.

A fascinating thing happened in Y4: futures prices were at their PF and PI equilibrium value and a particular type I outsider began to mimic the trading strategy of the type I insider. The insider sold two futures in PA, and the outsider sold three; the insider sold six futures in PB, and his "shadow" sold five. This mimicry is also evident in the spot transactions of these two traders (discussed below).

In Y5 there was a return to state of nature Y. Following some intraperiod leakage of insider information in PA spot trading there was heavy futures trading at the PF and PI equilibrium prices. Futures prices in PB remained at this equilibrium.

Period B (Spot)

Prices opened at a low level in Y1, generally at the PI value. All of the buyers were insiders and all of the sellers were outsiders. In PB spot prices were midway between the PI and PF equilibria; recall that initial PB futures price signals understated the true market value of a certificate in PC.

In Y3 an excellent futures price signal was available but there was only one spot trade in PB. This is easily explained: type II is predicted by PF equilibrium to hold spot certificates in PA *and* PB and, by the end of PA in Y3, indeed held all 18 spot certificates (note that type II is also predicted to hold spot certificates in PA by the EPI equilibrium concept). Thus there was simply no need for any spot trade in PB. The same result is observed in Y5, another Y year.

Y4 is an X year, with excellent futures price signals. Recall that a certain outsider had identified the insider of "his" type; this accounts for their parallel spot purchases in PB. Despite this partial leakage of insider information to one outsider, PB spot prices were kept quite low at their PI equilibrium value

Spot allocations in the Y years Y1, Y3, and Y5 were consistent with the PF prediction: 17/18, 18/18, and 18/18, respectively. The spot allocations in the X years Y1 and Y4 are severely misallocated relative to the PF prediction, with a mere 3/18 and 1/18 being correctly predicted. Moreover, only 8/15 and 1/17 of these deviations are explained by the alternative PI prediction. The explanation is the low PB spot prices (relative to the PF value) and the ability to lock in profits with futures trades. Given the prevailing low PB spot prices and the existence of a liquid futures contract, the erstwhile type I "culprits" did not lose on their spot purchases. This explanation accounts for 7/15 and 16/17 of the PF deviations in Y1 and Y4. Given that type III agents were not buying in PB, there was no opportunity cost to the type I agents pursuing their strategy. Note also that there was only a $0.05 difference in the PB dividends (per certificate) of types III and I.

Period A (Spot)

In Y1 spot prices for PA certificates opened below the PI value, moved quickly to the EPI value, and settled around the PF value. The type I insider was an early bidder and bought one unit at $1.10 and then another unit at $1.20; he thereby leaked this higher valuation to the market. The selling insiders generally held off until prices were above $1.25, reinforcing the leakage effect by refusing to supply at "low" prices. Note also the tendency for spot prices to reach the PF value despite traders' not having seen any PB spot valuations. This may have been because two traders were able to sell PC futures at the "overvalued" PC price (two of the four futures certificates were purchased by the type I insider), which provided some justification for "high" PA prices.

In Y2 there was a change in the state of nature. Spot prices opened below the URE value, passed through that value, and converged to the PI equilibrium. The final spot allocations hide heavy early buying by (non-PF-predicted) insiders at low prices for resale at higher prices later in the p
Spot prices in Y3, another Y year, opened where they had closed
(namely, at the PI value). There was then a sustained increase to
equilibrium value, accompanied by a parallel increase in futures prices
their URE value to their PF equilibrium. The type II insider was an
buyer at the lower prices.

Prices in Y4, the second X year, opened below the PI equilibrium for
year and settled just above that value. Despite good futures price signals
type I insider and his outsider "shadow" were able to buy spot without any apparent leakage.

In Y5 spot prices picked up where they had closed in Y3, the most recent Y year. After opening midway between the PI and PF equilibria for a Y year they increased steadily toward the PF equilibrium. The type II insider bought six of the first seven spot certificates traded, clearly revealing his inside information. The first five spot sales were by outsiders; again, the selling insiders held off before trading their endowments, in the knowledge that spot prices should be higher than the price that was offered initially.

Spot allocations in the X years Y1 and Y4 improve from 6/18 to 13/18 relative to the PF prediction, with the competing EPI prediction explaining 7/12 and 3/5 of the PF misallocations. Allocations in the Y years, Y2, Y3, and Y5, are consistent with the PF prediction: 10/18, 18/18, and 18/18, respectively, are correctly predicted.

Appendix C

Additional Results

TABLE A1

CERTIFICATE ALLOCATIONS PERIOD C AND FUTURES

EXPERIMENT AND YEAR	STATE	FUTURES CERTIFICATES						PERIOD C SPOT CERTIFICATES		
		Period A			Period B					
		I	II	III	I	II	III	I	II	III
1:										
1	...	...	...	...	...	...		4	0	14*/**
2	...		...	...	...	...	...	0	0	18*/**
3	...	...	...	...	...	...	...	0	2	16*/**
4	...	...	...	...	...	...	...	0	0	18*/**
5	...	...	...	...	...	...	...	...	0	18*/**
2:										
1	...	0	−3	3*	−1	−13	14*/**	...	...	...
2	...	0	−5	5*	0	−18	18*/**	...	...	...
3	...	0	−5	5*	0	−18	18*/**	...	...	...
4		−7	0	7*	−2	−11	13*/**	...	...	...
5		−9	−4	13*	0	−15	15*/**	...	...	...

3										
1	X	...	...	...	...	...	...	0	4	14*/**
2	Y	...	...	...	...	...	...	18*/**	0	0
3	X	...	...	...	...	...	...	0	0	18*/**
4	Y	...	...	...	...	...	...	18*/**	0	0
5	X	...	...	...	...	...	...	5	2	11*/**
4.										
1	X	−1	−5	6*	−1	−12	13*/**	...	...	...
2	Y	0	2	−2*	13*/**	−12	−1	...	...	...
3	Y	−1	3	−2*	16*/**	−16	0	...	...	...
4	X	−7	4	3*	−1	−10	11*/**	...	...	...
5	X	0	0	0*	0	−17	17*/**	...	...	...
5.										
1	Y	...	...	...	...	...	...	0	0	18*/**
2	Y	...	...	...	...	...	...	0	2	16*/**
3	X	...	...	...	...	...	...	17*/**	1	0
4	X	...	...	...	...	...	...	16*/**	2	0
5	X	...	...	...	...	...	...	16*/**	2	0
6:										
1	X	−2	−2	4*	−7	7*/**	0	...	...	...
2	Y	0	−6	6*	−1	−15	16*/**	...	...	...
3	Y	0	−13	13*	0	−16	16*/**	...	...	...
4	X	−5	3	2*	−16	17*/**	−1	...	...	...
5	Y	3	−15	12*	0	−18	18*/**	...	...	...

NOTES.—A single asterisk indicates the weaker-form prediction and a double asterisk the stronger-form prediction. The trader type predicted to be long in futures is indicated. See table 3 above for a list of these predictions.

TABLE A2

Number of Certificates Misallocated Relative to Equilibrium Predictions

Experiment and Year	State	EPI	Period A PI	Period A URE	Period A PF	Period B PI	Period B PF
1.							
1	...	...	14	...	4	4	15
2			18	...	0	17	3
3		...	18	...	0	18	2
4	...	...	18	...	0	18	0
5		...	18	...	0	18	0
2							
1		...	5	...	5	5	5
2	...	...	5	...	5	0	0
3	...	...	5	...	5	0	0
4			0	...	0	7	7
5			4		4	2	2
3							
1	X	8	...	13	...	17	1
2	Y	14		4	...	5	13
3	X	10		8		14	4
4	Y	11	...	7	...	18	3
5	X	18	...	0		18	0
4							
1	X	13		12		16	5
2	Y	13		7	...	13	6
3	Y	18	...	9		16	2
4	X	18		8	...	18	7
5	X	18		10	...	18	1
5.							
1	Y	6	6	12	0	1	17
2	Y	0	0	18	0	18	0
3	X	14	4	4	4	4	12
4	X	7	11	11	11	2	4
5	X	16	15	15	15	0	2
6.							
1	X	11	12	12	12	10	15
2	Y	10	10	18	8	18	1
3	Y	18	18	18	0	18	0
4	X	15	5	5	5	17	17
5	Y	18	18	18	0	18	0

Source.—Derived from table 4 above

TABLE A3

Coefficient of Variation

Experiment and Year	State	Period A	Period B	Period C
1:				
1	. .	066	061	.175
2	. .	042	010	048
3	. . .	031	010	030
4	. .	018	.027	015
5	. . .	.023	.003	004
2:				
1	. . .	094	093	001/ 125
2	.	.036	.017	.013/.013
3	.	043	008	001/ 005
4	.	.005	001	013/ 004
5	. .	.015	.016	.013/ 012
3:				
1	X	.092	.029	021
2	Y	.010	.018	040
3	X	006	002	015
4	Y	013	018	013
5	X	022	009	0
4:				
1	X	052	013	022/ 013
2	Y	009	010	001 003
3	Y	.008	006	008/ 005
4	X	008	003	005/ 013
5	X	.003	004	/ 005
5:				
1	Y	028	015	017
2	Y	074	013	[illegible]
3	X	035	011	
4	X	017	032	
5	X	022	019	
6.				
1	X	.060	125	[illegible]
2	Y	038	016	
3	Y	029		[illegible]
4	X	035	035	[illegible]
5	Y	018		[illegible]

Source.—Derived from figs. 1–6 above.

TABLE A4

Proportional Improvement in Coefficient of Variation

Experiment	State	Period A	Period B
1	. .	348	049
3	X	239	310
	Y	1 3	1 0
5	X	629	1 727
	Y	2 643	867
2	. . .	159	172
4	X	058	308
	Y	889	6
6	X	583	280
	Y	474	

Note.—The measure shown is the CV in the terminal year divided by the CV in the initial year; the CV measures are listed in table A3.

TABLE A5

EFFECT OF FUTURES MARKETS ON SPOT-PRICE VARIABILITY FPP*b* RESULTS

EXPERIMENT	SPOT OR FUTURES	COEFFICIENT OF VARIATION BY TRADING PERIOD							PROPORTIONAL CV IMPROVEMENT
		1	2	3	4	5	6	7	
1	Futures	1682	0713	1040	0414	0205	0252	0079	0469
3	Futures	.2355	1105	0235	.0	.0099	0	.0034	.0144
5	Futures	0418	.1115	0332	.0440	.0667	0232	.0131	3134
6	Futures	0708	.0869	.0638	.0719	0610	0496	.0215	3037
8	Futures	.0	0180	0878	.0249	.0401	0352	.0374	2 0780
2	Spot	0608	0136	0033	.0079	0105	0026	.0456	7500
4	Spot	.0263	.0331	0230	0740	0267	.0290	0221	.8403
7	Spot	1162	0245	0185	0084	.0146	0083	0132	1136
9	Spot	.0379	0076	.0163	.0104	0	0027	.0035	0923
Probability of favorable effect of futures market		.278	032	008	.365	.095	.206	794	.722
Pooled probability					.003				

NOTE —CV derived from data in table 4 of Forsythe, Palfrey, and Plott (1982*b*) Proportional CV improvement is defined in table A4 (period 2 CV used instead of zero period 1 CV in experiment 8, this has no effect whatsoever on Mann-Whitney test results) The probabilities are obtained from a Mann-Whitney test as used in table 8 The "favorable effect" is a "reduced CV" and "increased proportional CV improvement," respectively

TABLE A6

IMPROVEMENT IN TRADING EFFICIENCY

Experiment	State	Measure
1	. .	-35
3	X	9
	Y	−4
5	X	28
	Y	−10
2	. .	5
4	X	−23
	Y	−4
6	X	−15
	Y	−39

NOTE.—Improvement measured as the change in the percent (stronger-form) forecast error between the initial and terminal year, a positive value indicates an increase in forecast errors (i e , forecast deterioration) and a negative value a decrease in forecast errors (i e , forecast improvement), see table 7 for original data

References

Arrow, Kenneth J , and Hahn, Frank. *General Competitive Analysis* San Francisco: Holden Day, 1971.

Bray, Margaret M. "Futures Trading, Rational Expectations, and the Efficient Markets Hypothesis " *Econometrica* 49 (May 1981): 575–96.

Cox, Charles C. "Futures Trading and Market Information." *J P E* 84 (December 1976): 1215–37.

Easley, David, and Ledyard, John O. "A Theory of Price Formation and Exchange in Oral Auctions." Discussion Paper no. 461. Evanston Northwestern Univ., Center Math. Studies Econ. and Managem 1983 (revised).

Fama, Eugene F. "Efficient Capital Markets: A Review of Theory and ical Work." *J. Finance* 25 (May 1970): 383–417.

Forsythe, Robert; Palfrey, Thomas R., and Plott, Charles R. "Asset in an Experimental Market." *Econometrica* 50 (May 1982): 537–67

———. "Futures Markets and Informational Efficiency: A Laboratory nation." GSIA Working Paper no. 11-82-83. Pittsburgh, Carnegie-Mellon Univ., 1982. (*b*) (Forthcoming in *J. Finance*)

Friedman, Daniel. "Price Formation in Double Auction Markets." Discussion Paper no. 278. Los Angeles: U.C.L.A , Dept Econ , 1982

———. "On the Efficiency of Experimental Double Auction Markets." *A E R.* 74 (March 1984): 60–72.

Friedman, Daniel, and Harrison, Glenn W. "Price Formation in Experimental Asset Markets: A Bayesian Approach." Unpublished manuscript. London. Univ. Western Ontario, Dept. Econ , 1984.

Friedman, Daniel; Harrison, Glenn W.; and Salmon, Jon W "The Informational Role of Futures Markets: Some Experimental Evidence." In *Futures Markets—Modelling, Managing and Monitoring Futures Trading*, edited by M. E. Streit. Oxford: Basil Blackwell, 1983.

Glover, L. "Microdynamic Price Determination." Doctoral thesis proposal Los Angeles: U.C.L.A., 1983.

Grether, David M., and Wilde, Louis L. "Consumer Choice and Information: New Experimental Evidence on the Information Overload Hypothesis." Soc. Sci. Working Paper no. 459. Pasadena: California Inst. Tech., Division Humanities and Soc. Sci., 1982.

Grossman, Sanford J. "On the Efficiency of Competitive Stock Markets Where Traders Have Diverse Information." *J. Finance* 31 (May 1976): 573–85.

——— "The Existence of Futures Markets, Noisy Rational Expectations and Informational Externalities." *Rev. Econ. Studies* 44 (October 1977): 431–49.

———. "Further Results on the Informational Efficiency of Competitive Stock Markets." *J. Econ. Theory* 18 (June 1978): 81–101.

Grossman, Sanford J., and Stiglitz, Joseph E. "Information and Competitive Price Systems." *A.E.R. Papers and Proc.* 66 (May 1976): 246–53.

———. "On the Impossibility of Informationally Efficient Markets." *A.E.R.* 70 (June 1980): 393–409.

Hardy, Charles O. *Risk and Risk-Bearing.* Rev. ed. Chicago: Univ. Chicago Press, 1940.

Harrison, Glenn W. "Experimental Futures Markets." In *Futures Markets: Their Establishment and Performance,* edited by B. A. Goss. London: Croom Helm, 1984.

Harrison, Glenn W.; Smith, Vernon L.; and Williams, A. W. "Learning Behavior in Experimental Auction Markets." Unpublished manuscript. London: Univ. Western Ontario, Dept. Econ., 1983.

Hoel, Paul G. *Introduction to Mathematical Statistics.* 4th ed. New York: Wiley, 1971.

Plott, Charles R., and Sunder, Shyam. "Efficiency of Experimental Security Markets with Insider Information: An Application of Rational-Expectations Models." *J.P.E.* 90 (August 1982): 663–98.

———. "Rational Expectations and the Aggregation of Diverse Information in Laboratory Security Markets." Soc. Sci. Working Paper no. 463. Pasadena: California Inst. Tech., Division Humanities and Soc. Sci., 1983.

Svensson, Lars E. O. "Sequences of Temporary Equilibria, Stationary Point Expectations, and Pareto Efficiency." *J. Econ. Theory* 13 (October 1976): 169–83.

Vickrey, William. "Counterspeculation, Auctions, and Competitive Sealed Tenders." *J. Finance* 16 (March 1961): 8–37.

Zellner, Arnold. *An Introduction to Bayesian Inference in Econometrics.* New York: Wiley, 1971.

Futures Markets and Production Decisions

Alan J. Marcus

Boston University

David M. Modest

Columbia University

This paper investigates the use of futures prices in making production decisions. We derive a preference-independent production rule for firms that face both demand and production uncertainty. This rule is compared to a simple "marginal cost equals future price" rule, which previously has been suggested for firms with deterministic output. Data for agricultural producers are used to examine the importance of output uncertainty in the determination of the proper production rule. Our analysis suggests that for many crops the simple rule is sufficiently accurate to be a useful guide to production.

I. Introduction

The potential use of futures markets in a firm's production dec
a subject of current interest. Several authors have made the intu
plausible argument that futures prices provide information th
be used to guide production, storage, and processing decisions (see, e.g., Black 1976), and this argument has been formalized in a variety of models. Some of these models suggest that production decisions can be based solely on futures prices, regardless of subjective beliefs

---

We thank Barry Goldman, Bob McDonald, Krishna Ramaswamy, and Mahadevan Sundaresan for helpful discussions, and Tom Earley of Schnittker Associates for data. Aid from the Center for the Study of Futures Markets, Columbia University, and Boston University School of Management is gratefully acknowledged.

[*Journal of Political Economy*, 1984, vol. 92, no. 3]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0010$01.50

about the probability distribution of future spot prices and regardless of attitudes toward risk. In particular, several papers have shown that when production is nonstochastic, the optimal production rule is to set marginal cost equal to the discounted futures price. In these models, expectations about future spot prices affect only the firm's hedge position in futures contracts. This result appears, for instance, in Danthine (1978), Holthausen (1979), Anderson and Danthine (1980, 1981), and Feder, Just, and Schmitz (1980).

However, for many firms production uncertainty is a significant concern. When output is stochastic, the appealing generalization of the textbook rule for profit maximization—price equals marginal cost—breaks down. Anderson and Danthine (1980) derive the optimal production rule for risk-averse firms with mean-variance utility functions and find that it involves the producer's risk aversion as well as the expected future spot price and the expected marginal product. Thus, effective separation of ownership and control is difficult because the optimal rule is a function of individual preferences and beliefs. Similar preference-dependent results have been derived in the classic papers of Sandmo (1971) and Leland (1972), who studied the production decisions of firms facing uncertain demand in the absence of futures markets. As far as we know, no production rules for the stochastic output problem independent of individual preferences have appeared in the literature.

In this paper, we derive a preference-independent production rule.[1] We focus on the problem of agricultural producers for whom production uncertainty seems especially important and for whom active futures markets already exist. However, our production rule in principle is applicable to other firms in competitive industries that face production uncertainty.

We have found that the continuous-time framework used in the options-pricing literature greatly simplifies the analysis of optimal production decisions. The simplification arises because the value of a farmer's crop, once planted, and the futures price for that commodity are affected by the same stochastic state variables, such as weather or demand fluctuation. In general, the particular functional forms of the relationships will be quite different. However, if each of the state variables follows a Wiener diffusion process, then we show that by mixing appropriate numbers of futures contracts with its production position in the crop, an agricultural firm can render its total return free of all systematic risk, with only risk unique to that firm remaining.[2] Because the unique risk is diversifiable, it follows that the value-maximizing production rule for any publicly held firm is independent of attitudes toward risk and therefore allows for separation of ownership and management.[3]

---

[1] Our rule is independent of preferences in that the production decision does not depend on the managers' or owners' attitudes toward risk. It will, however, depend on the consumers' preferences through the demand functions.

In Section II we present the model and derive closed-form solutions for optimal production rules when there is both price and quantity uncertainty. Given the simplicity of the rule that "marginal cost equals the discounted futures price," it is of some interest to compare that rule with the more general one derived in this paper. We provide analytic comparisons of the two rules in Section II. In Section III we use data for agricultural producers to measure the importance of output uncertainty in the determination of the proper production rule. We find that for most crops that are not subject to agricultural price supports, the simple rule that marginal cost equals futures price is sufficiently accurate to be a useful guide to production. However, this is not the case for all crops, especially those subject to price supports. Section IV is the conclusion.

II. Production under Demand and Output Uncertainty

A. *Market Setting*

This section considers the production decision of an agricultural producer. Firms (farmers) are assumed to be perfectly competitive pric takers. Further, there is assumed to be a well-developed futu ket and riskless borrowing and lending at the constant r production setting is taken to be point input–point outpu is planted at time zero and harvested and sold at time T production decisions are allowed. However, we allow capit to be open at all times so that portfolios can be contin balanced. The crop is assumed to be a normal good in that demanded rises with increases in wealth or income and declines wi the market price, P, rises. We assume a constant elasticity functional form. The dependence of demand on wealth or income occurs through a shift parameter, S, which can be thought of as measured by a stock market index.[4] Market demand for the good, Q^d, at the time of the harvest, T, is then given by

---

[2] The capital asset pricing model (CAPM) provides a traditional rationale for ignoring diversifiable risk. However, the irrelevance of diversifiable risk is a more fundamental property. Ross (1975) provides a lucid discussion of this point. It is shown below that if demand is also uncertain then another asset is needed to construct the hedged portfolio.

[3] Privately held firms in which owners have invested large fractions of their personal wealth (and hence are not well diversified) will not be indifferent to unique risk. However, output in the U.S. economy is dominated by publicly held firms whose owners can easily diversify away unique risk, and even in the agricultural sector an increasing share of output is accounted for by corporate farms.

$$Q^d(P, S, T) = kS(T)^{\beta}P(T)^{-\epsilon}, \quad \beta > 0, \epsilon > 0, \tag{1}$$

where $P(T)$ and $S(T)$ are random variables as of time $t < T$, that is, prior to the harvest, and β and ϵ are the wealth and price elasticities of demand. The supply of the good available at T, $Q^s(T)$, will equal the initial stock of the good, $Q^s(0)$, plus output over the period. For simplicity, we assume that the initial stock is zero.[5] Finally, we let $Q(t)$ denote the time t expectation of $Q(T)$. Units of output are chosen such that the scaling parameter, k, can be set equal to one.

In the absence of price supports, the market would clear at price

$$P(T) = S(T)^{\beta/\epsilon}Q(T)^{-1/\epsilon}, \tag{2}$$

which would provide total revenue to all farmers of

$$V[S(T), P(T)] = P(T)Q(T) = S(T)^{\beta/\epsilon}Q(T)^{1-(1/\epsilon)} = S(T)^{\beta}P(T)^{1-\epsilon}. \tag{3}$$

Indexing producers by j, we assume that the share of any producer in total output is given by $\theta_j(T)$, which is a random variable for times $t < T$. This implies that the time T revenue to each farmer equals

$$V_j[\theta_j, S(T), P(T)] = \theta_j(T)S(T)^{\beta}P(T)^{1-\epsilon}, \tag{4}$$

where $1 \gg \theta_j(T) \geqslant 0$, and $\Sigma_j\, \theta_j(T) = 1$.

The condition that the market shares sum to one places some restrictions on the covariance structure of the θ_j. In the agricultural context, for example, one would expect the changes in market shares for two contiguous farms to be highly correlated, while the correlation of either's market share with a farm in another weather system would likely be negative. However, these restrictions will not be important in a perfectly competitive industry with many producers $(\theta_j[T] \ll 1)$.

The perfect competition assumption places further restrictions on θ_j. These are that changes in θ_j, which affect only the distribution of output among firms, are uncorrelated with economy-wide variables such as aggregate demand and supply. In particular, we assume θ_j is uncorrelated with spot and futures market prices or with any systematic factors affecting aggregate demand, such as S. This last assumption is reasonable only under perfect competition, where producers would not perceive a correlation between their market share and aggregate output. Such a correlation would be perceived by a large producer for whom a positive shock to production simultaneously raises the expected values of both θ_j and aggregate supply, Q.

---

[4] The analysis can also be extended to allow for other systematic effects on demand, besides that coming from S.

[5] The analysis could be easily extended to incorporate nonzero inventories. However, incorporating inventories into the analysis adds no further insights to the problem at hand, and therefore we ignore them.

We assume that the goal of the producer is to maximize the value of the firm. The value-maximization strategy is justified in Merton (1982) under the assumption that no individual producer has any effect on the ex ante distribution of returns available to investors (i.e., the efficient frontier is independent of production decisions). The restriction that the θ_j are uncorrelated with all economy-wide variables is sufficient to satisfy this condition.

The farmer's (firm's) decision rule is then conceptually simple: Increase the size of the crop as long as the increment to ex ante value exceeds the marginal cost of planting. While the value of the crop and hence the level of planting will depend on expectations about $S(T)$ and $P(T)$, we assume that once the crop is planted the farmer has little control over final output. Yield is primarily determined by exogenous factors, such as weather, and is not affected by supply responses to changes in the market or futures prices.

The perfect competition assumption implies that each farmer acts as if his decisions affected neither aggregate output nor the aggregate value of the crop. More formally, let V = the current value of the claim to the stochastic revenue from the aggregate crop (measured in dollars), $\theta_j(I_j)$ = the fraction of the crop at time T produced farmer j, where T has been dropped as an argument for simplicity, and I_j = the value of the factor inputs of farm sured in dollars). Under perfect competition, each pro ceives that[6]

$$\frac{\partial\theta_j}{\partial I_j} \equiv \theta_j'(I_j) > 0, \quad \frac{\partial^2\theta_j}{\partial I_j^2} < 0, \quad \frac{\partial V}{\partial\theta_j} = 0.$$

It follows that the value of any individual farmer's crop at time T is given by $V_j(T) = \theta_j(I_j)V(T)$, and that each producer solves the following problem at time 0:

$$\max_{I_j} V_j(0) - I_j. \tag{5}$$

In a perfectly competitive industry where each producer perceives his market share to increase with his input level conditional on the input levels of the other producers, value maximization implies that the farmer should increase the size of the crop as long as the increment to value, $\theta_j'(I_j)V$, exceeds the marginal cost of planting. Once a technique for valuing the crop is given, this is a general production rule. We show below in Section IIC that when output is nonstochastic this rule reduces to setting marginal cost equal to price.

---

[6] The first two assumptions guarantee that the production rule is well defined.

B. Valuation of the Crop

The current value of the crop can be derived using the techniques of contingent claims analysis developed in Black and Scholes (1973), Merton (1973, 1977), Constantinides (1978), and Fischer (1978). By combining a claim on the revenues from the crop, futures contracts on the commodity, equity holdings of the stock index, S, and riskless borrowing or lending, the farmer can eliminate systematic risk.[7] The equilibrium expected rate of return on this hedged portfolio must be the riskless rate, r, because the portfolio is subject only to diversifiable risk. We show below that this equilibrium condition can be used to specify a partial differential equation that can be solved for the value of the crop.

The futures price is the price at which one can contract today to buy or sell a commodity at some fixed future date. Futures contracts are marked to market and new equilibrium futures prices are established daily. Assume that the dynamics of the futures price, F,[8] and the stock

---

[7] This is a direct consequence of modeling uncertainty by diffusion processes whose continuous sample paths enable farmers to eliminate systematic risk by continuously rebalancing their portfolios of futures contracts and the stock index. Thus, the preference-free production rules derived below follow directly from the continuous-time nature of the model. This ability to rebalance continuously is missing from discrete-time models, like that of Anderson and Danthine (1980), although Rubinstein (1976) shows that the Black-Scholes formula can be derived when investors are able to trade only at discrete intervals provided that the joint distribution of stock prices and marginal utility is log normal.

[8] It might seem more natural to treat quantity, Q, and S as the state variables of the system. F would then be solved for endogenously. However, it is mathematically equivalent and operationally simpler to describe the market equilibrium in terms of F and S, because F is observable at $t < T$. This facilitates empirical estimation of the valuation formula. Treating Q and S as the state variables and assuming the government does not actively intervene to support the spot price, we can show that the equilibrium futures price is given by

$$F(t) = Q(t)^{\tilde{\delta}} S(t)^{\tilde{\beta}} \exp\left[\tilde{\beta} r + \tilde{\delta}(\tilde{\delta} - 1)\frac{\sigma_Q^2}{2} + \tilde{\beta}(\tilde{\beta} - 1)\frac{\sigma_S^2}{2} + \tilde{\delta}\tilde{\beta}\sigma_Q\sigma_S\rho_{SQ}\right](T - t),$$

where $\tilde{\delta} \equiv -1/\epsilon$ and $\tilde{\beta} \equiv \beta/\epsilon$. Using Ito's lemma we can then show that the equilibrium dynamics for the futures price are given by (6a) where

$$\begin{aligned} \alpha_F &= \tilde{\beta}(\alpha_S - r) + \tilde{\delta}\alpha_Q, \\ \sigma_F dz_F &= \tilde{\beta}\sigma_S dz_S + \tilde{\delta}\sigma_Q dz_Q. \end{aligned} \qquad (6a')$$

index, S, are described by diffusion processes:[9]

$$\frac{dF}{F} = \alpha_F dt + \sigma_F dz_F, \tag{6a}$$

$$\frac{dS}{S} = \alpha_S dt + \sigma_S dz_S, \tag{6b}$$

where α and σ^2 are instantaneous expected growth rates and variances rates, respectively. The vibration of actual changes around expected changes is modeled by the Wiener processes dz_F and dz_S; these processes have zero mean, unit variance, and an instantaneous correlation of ρ. Because futures contracts are settled daily, the value of any contract is zero. Further, when the contract matures, $F(T)$ will equal $P(T)$.[10]

Let $V[\theta(t), F(t), S(t), t]$ represent the value at time t of the claim to a farmer's revenue, where subscripts j have been dropped for notational simplicity. $F(t)$ and $S(t)$ are, respectively, the futures price and the value of the stock market index at date t; $\theta(t)$ is the expectation at time t of the market share that the producer will realize at time T. Expectations are assumed rational in that the expected value of the change in θ is zero. Actual expectations will change only as new (unanticipated) information becomes available. We therefore assume that the dynamics of θ can be described by the diffusion process

$$\frac{d\theta}{\theta} = \sigma_\theta dz_\theta,$$

which has zero drift. It follows from equation (6c) that the
of $\theta(T)$ at date $t < T$ is simply $\theta(t)$—a consequence of the
of rational expectations. Using the dynamics given in (
show in the Appendix by standard methods from the opt
literature that the valuation function, $V(\cdot)$, must satisfy
differential equation

$$\begin{aligned} &\tfrac{1}{2}V_{\theta\theta}\theta^2\sigma_\theta^2 + \tfrac{1}{2}V_{SS}S^2\sigma_S^2 + \tfrac{1}{2}V_{FF}F^2\sigma_F^2 \\ &+ V_{SF}SF\sigma_S\sigma_F\rho + V_t - r(V - V_S S) = 0. \end{aligned} \tag{7}$$

---

[9] This assumption is not restrictive in that this approach can also be used to derive valuation formulae in the presence of mixed jump and diffusion processes following the analysis of Cox and Ross (1976) and Merton (1976). However, a closed-form characterization of the solution is then no longer available. We also assume, for convenience, in eq. (6) that σ is time invariant. Samuelson (1965) and Anderson (1982), among others, discuss the merits of this assumption.

[10] In practice, the price any individual farmer receives at T per unit of output need not converge to $F(T)$ because of differences in the quality of the crop or the delivery location. We ignore this "basis risk," although under the assumption that basis risk is unsystematic, our valuation functions would still hold with only minor modifications.

The value of the claim to the output, $V(\theta, S, F, t)$, is given by the solution for V of equation (7) with the terminal condition

$$\begin{aligned} V[\theta(T), F(T), S(T), T] &= \theta(T)P(T)Q(T) \\ &= \theta(T)F(T)Q(T) = \theta(T)S(T)^{\beta}F(T)^{1-\epsilon}. \end{aligned} \tag{8}$$

The valuation function that satisfies (7) and (8) is:

$$\begin{aligned} V[\theta(t), F(t), S(t), t] = \theta(t)F(t)^{1-\epsilon}S(t)^{\beta} \exp\{[-r + \beta r + \tfrac{1}{2}\epsilon(\epsilon - 1)\sigma_F^2 \\ + \tfrac{1}{2}\beta(\beta - 1)\sigma_S^2 + \beta(1 - \epsilon)\rho\sigma_S\sigma_F](T - t)\}. \end{aligned} \tag{9}$$

In a risk-neutral world the valuation function (9) would have a simple interpretation—that the value of the output is the present discounted value (at the riskless rate) of expected revenues, since expected revenues are given by

$$\begin{aligned} \theta(t)F(t)^{1-\epsilon}S(t)^{\beta} \exp\{[\beta\alpha_S + \tfrac{1}{2}\epsilon(\epsilon - 1)\sigma_F^2 \\ + \tfrac{1}{2}\beta(\beta - 1)\sigma_S^2 + \beta(1 - \epsilon)\rho\sigma_S\sigma_F](T - t)\}, \end{aligned}$$

and in a risk-neutral equilibrium α_S would equal r.

The value of the crop is larger (i) as the expected market share rises: $\partial V_j(\cdot)/\partial\theta_j = V > 0$; (ii) with increases in the futures price for $0 < \epsilon < 1$ and with decreases in the futures price for $\epsilon > 1$, $\partial V_j(\cdot)/\partial F(t) = (1 - \epsilon)(V/F)$; (iii) with increases in wealth, $\partial V_j(\cdot)/\partial S(t) = \beta(V/S) > 0$; and (iv) with decreases in the interest rate $\partial V_j(\cdot)/\partial r = -(T - t)(1 - \beta)V < 0$. The effects of changes in the demand elasticities, β and ϵ, and the variance of the state variables, σ_S and σ_F, on the value of the crop are, in general, ambiguous. However, for $\rho > 0$ and $0 < \epsilon < \tfrac{1}{2}$, as demand becomes less price inelastic (larger ϵ) the value of the crop decreases, and for $\beta > \tfrac{1}{2}$, as demand becomes less wealth inelastic (larger β) the value of the crop increases. For $\beta > 1$ an increase in the variance of percentage changes in wealth (larger σ_S) unambiguously increases the value of the crop.

The valuation function (9) can be used to infer the relative proportions of the crop, the stock index, and the futures contracts held in the hedged portfolio. The portfolio consists of V_F short positions in the futures contracts and V_S short positions in the stock market index. Therefore, the ratio of the dollar values of the holdings of each security to the value of the crop are FV_F/V and SV_S/V. SV_S gives the dollar value of the stock market holdings in the hedged portfolio, while FV_F gives the dollar value of the amount that the farmer will receive for delivery of V_F futures contracts. From expression (9) it follows that $FV_F/V = 1 - \epsilon$ and $SV_S/V = \beta$. It is interesting to note that for $\epsilon > 1$, the farmer will take a long position in the futures market.[11] This is because for a price elasticity greater than one, revenue from the crop varies inversely with the spot price. Thus the appropriate hedge involves a long position in futures contracts for which profits increase with the spot price. It should also be noted that the weights in the hedge portfolio of the futures contract and stock market index will change continuously with the passage of time and fluctuations in F and S. The producer must continuously reevaluate V_S and V_F to rebalance the hedge portfolio. This requires knowledge of F and S at all points in time, as well as of the demand elasticities.

Notice that the contingent claims technique is extremely flexible. For example, when agricultural price supports are implemented at a support price $\overline{P}$, the appropriate terminal condition is given by

$$V[\theta(T), F(T), S(T), T] = \max\,[\theta(T)S(T)^{\beta}F(T)^{1-\epsilon},\ \theta(T)S(T)^{\beta}F(T)^{-\epsilon}\overline{P}]. \tag{10}$$

We have assumed in (10) that the government lets the market clear and compensates the farmer for the difference between the market and support prices. With the exception of dairy products, all major commodities are supported in this way (Congressional Budget Office 1976). Therefore, once the crop is planted, the support system will affect neither spot nor futures prices. The hedge portfolio still can be constructed by combining positions in the crop, futures contracts, the stock market, and riskless borrowing or lending. Therefore the
of any firm's future revenues still must satisfy equation (7)
ing over j, the valuation formula that satisfies (7) and (10

$$V(F, S, t) = V_0N(D_1) + V_{\overline{P}}[1 - N(D_2)],$$

where $\tau = T - t$,

$$V_0 = e^{(g-r)\tau}(S^{\beta}F^{1-\epsilon}),$$

$$V_{\overline{P}} = e^{(m-r)\tau}(S^{\beta}F^{-\epsilon}\overline{P}),$$

$$g = \beta r + \tfrac{1}{2}\beta(\beta - 1)\sigma_S^2 - \tfrac{1}{2}\epsilon(1 - \epsilon)\sigma_F^2 + \beta(1 - \epsilon)\sigma_F\sigma_S\rho,$$

$$m = \beta r + \tfrac{1}{2}\beta(\beta - 1)\sigma_S^2 + \tfrac{1}{2}\epsilon(1 + \epsilon)\sigma_F^2 - \beta\epsilon\sigma_F\sigma_S\rho,$$

$N(\cdot)$ is the cumulative normal distribution,

$$D_1 = \frac{(V_0/V_{\overline{P}}) + (\sigma_F^2\tau/2)}{\sigma_F\sqrt{\tau}},$$

and

$$D_2 = D_1 - \sigma_F\sqrt{\tau}.$$

---

[11] Estimates of ϵ for 11 different commodities are provided in Sec. III (table 1). The vealth elasticity of demand, β, for agricultural goods is expected to be well below one-ialf

As would be expected, an increase in the support level, $\bar{P}$, increases the value of the claim, as does an increase in wealth, S.[12] Equation (11) can be interpreted as a modified Black-Scholes formula for a put option with a randomly evolving exercise price $V_{\bar{P}}$ on an asset with value V_0. V_0 is, from equation (9), the present value of the crop in the absence of supports. Similarly, $V_{\bar{P}}$ is the present value of the crop if priced at $\bar{P}$.

C. Optimal Production Rules

1. General Rule

The farmer chooses the size of the crop so as to maximize the value of his harvest. From expression (5) it follows that the optimal choice for I_j (the value of the factor inputs) is implicitly defined by the first-order condition for value maximization:

$$\frac{\partial \theta_j(I_j)}{\partial I_j} V(\cdot) = 1. \tag{12}$$

Equation (12) simply states that when individual production decisions are optimal, one extra dollar of investment at the margin must lead to a one-dollar increase in the ex ante value of the crop, for every farm j. It follows that if any individual firm overestimates $V(\cdot)$, it will be incurring too high a marginal cost and will be "overinvesting" in the crop.

2. Special Case: Nonstochastic Output

What would happen if the farmer treated output as nonstochastic, as has much of the literature? In this case, the standard deviation of output (σ_Q in [6a'], n. 8) is mistakenly assumed to be equal to zero. The "incorrect" value function can then be derived by proceeding as above but imposing the false restriction that $\sigma_Q = 0$. Under this restriction, F and S are locally perfectly correlated, that is, $\rho = 1$. Further, expression (6a') in note 8 can be used to relate the instantaneous variances of F and S:

$$\sigma_F^2 = \frac{\beta^2 \sigma_S^2}{\epsilon^2}. \tag{13}$$

---

[12] The hedge ratio SV_S/V, as above, is equal to β. However, the hedge ratio FV_F/V is now more complicated and is equal to

$$\left\{V_0\left[(1-\epsilon)N(D_1) + \frac{e^{(g-m)\tau}n(D_1)}{\bar{P}\sigma_F\sqrt{\tau}}\right] + V_{\bar{P}}\left[-\epsilon + \epsilon N(D_2) + \frac{e^{(g-m)\tau}n(D_2)}{\bar{P}\sigma_F\sqrt{\tau}}\right]\right\} \Big/ V.$$

Imposing (13) in (9) gives the "incorrect" value function—the value of the crop if output were treated as nonstochastic (in the absence of price supports):

$$V[F(t), S(t), t|\sigma_Q = 0] = F(t)^{1-\epsilon}S(t)^{\beta} \exp\left\{\left[-r + \beta r + \frac{\beta\sigma_S^2}{2}\left(\frac{\beta}{\epsilon} - 1\right)\right](T - t)\right\}, \tag{14}$$

where θ has been dropped for notational simplicity. It is now straightforward to show that the improper use of (14) to value the crop leads to the standard rule that marginal cost equals the discounted futures price. When $\sigma_Q = 0$, it follows from equations (1) and (6) that:

$$E[Q(T)] = Q(T) = F(t)^{-\epsilon}S(t)^{\beta} \exp\left\{\left[\beta r + \frac{\beta\sigma_S^2}{2}\left(\frac{\beta}{\epsilon} - 1\right)\right](T - t)\right\}. \tag{15}$$

Thus we may rewrite (14) as

$$V(F, S, t|\sigma_Q = 0) = Q(T)F(t)e^{-r(T-t)}. \tag{16}$$

Inserting (16) into (12) and noting that $1/[\theta_j'(I_j)Q(T)]$ is the cost of one unit of output, we can rewrite condition (12) t the optimal production rule, when output is nonstochast the marginal cost of one unit of production equal to the futures price:

$$F(t)e^{-r(T-t)} = \frac{1}{[\theta_j'(I_j)Q]}$$

This is the standard result found in the literature. Thus our production rule includes the nonstochastic production rule as a special case.

D. Analytical Comparison of Production Rules

The application of (12′) when output is in fact stochastic will lead farmers to expend an inappropriate level of resources. As shown in equation (12), the optimal level of marginal cost is proportional to the farmer's perception of the ex ante value of the crop. Therefore, by comparing the estimates of $V(F, S, t)$ when output is treated as nonstochastic to its correct value, one may evaluate the percentage difference in marginal costs chosen by a farmer who uses the rule that marginal cost equals the discounted futures price instead of the correct rule:[13]

$$\frac{V[F(t), S(t), t|\sigma_Q = 0] - V[F(t), S(t), t]}{V[F(t), S(t), t]}, \tag{17}$$

where $V[F(t), S(t), t]$ is the solution to the partial differential equation given by (7) with the appropriate boundary condition.

1. In the Absence of Agricultural Price Supports

In the absence of price supports, the valuation formulae are given by (9) (stochastic output) and (14) (nonstochastic output). Substituting (9) and (14) into (17) yields the percentage difference in the two valuation formulae:

$$\frac{\exp\left\{\left[\frac{\beta\sigma_S^2}{2}\left(\frac{\beta}{\epsilon} - 1\right)\right](T - t)\right\}}{\exp\left\{\left[\frac{\epsilon}{2}(\epsilon - 1)\sigma_F^2 + \frac{\beta}{2}(\beta - 1)\sigma_S^2 + \beta(1 - \epsilon)\rho\sigma_S\sigma_F\right](T - t)\right\}} - 1. \tag{18a}$$

Empirical estimates of expression (18a) are presented below for 11 different commodities.

2. In the Presence of Agricultural Price Supports

In the presence of supports, the valuation rule with stochastic output is given by (11), while the assumption of nonstochastic output yields (14). Substituting these expressions into (17) yields the percentage difference in the marginal cost that the farmer would incur by following the simple rule that marginal cost equals the futures price (14) rather than the correct rule:[14]

---

[13] The percentage error in marginal cost is a measure of resource misallocation. It does not necessarily indicate the magnitude of welfare loss, which would more properly be measured by lost producer plus consumer surplus. However, that measure would require several inputs, such as farmers' production functions, which are not integral to the rest of the analysis. It also should be acknowledged that a small percentage error in marginal cost need not translate into a small error in output if the marginal cost curve is relatively flat. This again is an empirical issue beyond the scope of our analysis. However, we note that agricultural price support programs generally involve cropland set-aside provisions, which effectively limit the land allowable for planting. Limits on this input must induce a significant steepening of the marginal cost per unit of output.

$$\frac{\exp\left\{\left[\frac{\beta\sigma_S^2}{2}\left(\frac{\beta}{\epsilon}-1\right)\right](T-t)\right\}}{N(D_1)e^{(g-\beta r)}+N(D_2)e^{(m-\beta r)}\left(\frac{\bar{P}}{F}\right)}-1. \tag{18b}$$

Empirical estimates of (18b) are also presented below for 11 different commodities.

III. Production Rules for Various Commodities: Empirical Results

Equations (18a) and (18b) provide a comparison of the optimal values of $\theta_j'(I_j)$ under the simple versus the general rule of production. How good an approximation is the simple rule? This question can be answered given parameter estimates for the valuation formulae. Estimates of σ_F, σ_S, and ρ were taken from Bodie and Rosansky (1980), who examined returns on 23 commodity futures contracts, of which 11 were agricultural. Price elasticities, ϵ, were obtained from Schnittker Associates. These data are presented in table 1. Because we were unable to locate disaggregated wealth elasticity estimates, several values were examined. However, the results are quite insensitive to the exact choice of wealth elasticity, and therefore we prese only for a wealth elasticity of 0.25.[15]

Table 2 presents estimates of (18a) using the valuati derived for the no price support case. For all 11 commodi ered, the simple rule leads the farmer to produce too muc in most cases the error is quite small. The results presente are given for two alternative lengths of time until the harves ½ year and $T - t = 1$ year. The percentage errors range from percent to 11.7 percent, though only for one commodity, sugar, is the error greater than 5 percent. For all commodities considered the

---

[14] In the presence of agricultural price supports there is no natural "incorrect" rule to compare with the "correct" rule. Even if output were nonstochastic, farmers would not be expected to use the rule that discounted futures price equals marginal cost when $F < \bar{P}$, because the effective price must be at least $\bar{P}$. Thus a natural extension would be to assume the farmer sets the discounted value of max $(\bar{P}, F)$ equal to the futures price. An alternative "incorrect" rule could be derived by assuming that the farmer's only mistake is to assume incorrectly that $\sigma_Q = 0$ in (6a'). The valuation would then be given by imposing the variance restriction (13) on eq. (11). We have chosen the comparison above because it gives an upper bound to the maximum possible error.

[15] This number is consistent with previous estimates of aggregate income elasticities. See, e.g., Tomek (1977), who surveyed empirical studies of the demand for food.

TABLE 1

PARAMETER ESTIMATES

Commodity	σ_F	$\rho\sigma_F\sigma_S$	ϵ
Wheat	.308	−.0133	.10
Corn	.263	−.0154	.30
Oats	195	.0	.30
Soybeans	.323	−.0096	.40
Soybean oil	577	−.0233	65
Soybean meal	.356	.0086	.40
Potatoes	.421	−.0219	.15
Cotton	.362	−.0005	.20
Orange juice	.318	0042	.45
Cocoa	.546	−.0105	.40
Sugar	1.162	−.0863	15

SOURCES.—Bodie and Rosansky (1980, table 8), Schnittker Associates.

NOTE.—Bodie and Rosansky statistics were derived for 1950–76 using 3-month holding period yields. Estimates of $\rho\sigma_F\sigma_S$ derived by multiplying Bodie and Rosansky's estimates of CAPM β by estimates of the market (Standard & Poor's 500) variance (.0359).

percentage error is a nondecreasing function of the time to harvest and the wealth elasticity (results other than for $\beta = 0.25$ not shown). However, even for values of $\beta = 0.5$, only one commodity, sugar, has an error greater than 5 percent (16.4 percent for $T - t = 1$ year). Thus for all commodities except sugar the simple rule of setting marginal cost equal to the discounted futures price seems to be a very accurate approximation to the general rule.

TABLE 2

COMPARISON OF PRODUCTION RULES: NO PRICE SUPPORTS

	TIME TO HARVEST	
COMMODITY	6 Months	1 Year
Wheat	.0087	.0175
Corn	.0063	.0127
Oats	.0033	0066
Soybeans	0078	.0158
Soybean oil	.0020	.0413
Soybean meal	0078	.0157
Potatoes	0112	.0226
Cotton	.0076	.0152
Orange juice	.0067	.0134
Cocoa	.0197	.0398
Sugar	.0569	.1171

NOTE.—Estimates of expression (18a)

$$\frac{V[F(t), S(t), t|\sigma_Q = 0] - V[F(t), S(t), t]}{V[F(t), S(t), t]}$$

TABLE 3

COMPARISON OF PRODUCTION RULES: WITH PRICE SUPPORTS

	$\bar{P}/F$				
COMMODITY	.75	.90	1.0	1.1	1.25
Wheat	.0000	−.0336	−.0746	−.1263	−.2096
Corn	.0014	−.0277	−.0690	−.1233	−.2112
Oats	.0024	−.0144	−.0517	−.1089	−.2039
Soybeans	−.0041	−.0429	−.0867	−.1395	−.2226
Soybean oil	−.0561	−.1241	−.1745	−.2252	−.2973
Soybean meal	−.0088	−.0512	−.0954	−.1472	−.2278
Potatoes	−.0137	−.0598	−.1032	−.1522	−.2281
Cotton	−.0084	−.0490	−.0917	−.1421	−.2214
Orange juice	−.0047	−.0429	−.0867	−.1397	−.2231
Cocoa	−.0369	−.0968	−.1440	−.1931	−.2652
Sugar	−.1329	−.2012	−.2450	−.2865	−.3438

NOTE.—Time until harvest: 6 months. Estimates of expression (18b)

$$\frac{V[F(t), S(t), t|\sigma_Q = 0] - V[F(t), S(t), t]}{V[F(t), S(t), t]}$$

Tables 3 and 4 present estimates of (18b), which is the percentage difference in marginal cost that would result from using the simple "marginal cost equals futures price" rule instead of the optimal rule derived for stochastic output under price supports. Both tables present results for five possible ratios of support price to futures price. In almost all cases, following the simple rule would lead the produce too little. This is hardly surprising, since the g

TABLE 4

COMPARISONS OF PRODUCTION RULES: WITH PRICE SUPPORT

	$\bar{P}/F$				
COMMODITY	.75	.90	1.0	1.1	1.25
Wheat	−.0093	.0559	−.0992	−.1479	−.2232
Corn	−.0063	−.0504	−.0949	−.1464	.2262
Oats	.0006	−.0300	−.0714	.1249	−.2114
Soybeans	−.0205	−.0741	−.1205	−.1710	−.2467
Soybean oil	−.1246	−.2038	−.2550	−.3033	−.3689
Soybean meal	−.0298	−.0862	−.1326	−.1820	−.2553
Potatoes	−.0363	−.0934	.1380	.1845	.2534
Cotton	−.2069	−.0800	−.1243	−.1720	−.2438
Orange juice	−.0216	−.0748	−.1212	−.1718	−.2477
Cocoa	−.0808	−.1514	−.1998	−.2470	−.3131
Sugar	−.1992	−.2706	−.3136	.3530	.4059

NOTE.—Time until harvest: 1 year. Estimates of expression (18b)

$$\frac{V[F(t), S(t), t|\sigma_Q = 0] - V[F(t), S(t), t]}{V[F(t), S(t), t]}$$

guarantees farmers a minimum price that may be above the market price. As expected, the exact magnitude of the error is quite sensitive to the appropriate $\bar{P}/F$.

If the government were to guarantee farmers a minimum price equal to the current futures price (which can be interpreted as the risk-adjusted expected spot price), then for six of the 11 commodities the percentage error would be greater than 10 percent and for all the commodities the error would be greater than 5 percent. This assumes that the lag between planting and the harvest is 1 year. Over a 6-month horizon the simple rule would do slightly better, only three commodities having an error greater than 10 percent. However, for lower values of $\bar{P}/F$, such as 0.75, the error due to use of the simple rule is negligible.

Thus the accuracy of the simple production rule is affected by the existence and level of price supports. In practice, only a limited number of crops are supported (Congressional Budget Office 1976). As to the appropriate level of $\bar{P}/F$: for the 1982 harvest $\bar{P}/F$ was approximately 0.75 for oats and soybeans and approximately 1.0 for wheat, corn, and cotton. In all cases the error is less than 10 percent for a 6-month horizon and the simple rule is a very accurate approximation for these commodities.

To summarize, in the absence of price supports our results indicate that the simple rule is a good approximation to the more general one. In the presence of price supports we find similar results, as long as the support price is not set too high relative to the risk-adjusted expected spot price. Therefore, the futures price by itself, as suggested by Black (1976), does indeed provide valuable information for producers, storers, and users of commodities.

IV. Conclusions

This paper considers the production decision of a firm facing both demand and output uncertainty. In particular, we consider the case of an agricultural producer. We extend previous work on the theory of production under uncertainty to derive preference-free production rules even when output is stochastic. The production rules do depend on the existence of well-developed futures markets, which suggests that the development of markets that do not currently exist could lead to more efficient production decisions.

Appendix

Differentiating $V(\cdot)$ using Ito's lemma gives the dynamics for the change in the value of the claim:

$$dV = V_\theta d\theta + V_F dF + V_S dS + V_t dt + \tfrac{1}{2}V_{FF}(dF)^2 + \tfrac{1}{2}V_{SS}(dS)^2 + \tfrac{1}{2}V_{\theta\theta}(d\theta)^2 + V_{SF}\,dSdF + V_{F\theta}dFd\theta + V_{S\theta}dSd\theta. \tag{A1}$$

Notice that the last two terms in (A1) equal zero by the assumption that fluctuations in expected market shares, θ_j, involve purely idiosyncratic risk, which implies that the correlations between F and θ_j and S and θ_j are zero for all j.

Following Merton (1977), consider a portfolio consisting of (1) the claim to the farmer's revenue (value = V); (2) V_F short positions in the futures contract (value = 0); (3) V_S short positions in the stock index (value = $-V_S S$); and (4) borrowing of $V - SV_S$ at the riskless rate r. The instantaneous change in the value of such a portfolio will equal

$$dV - dF - dS - r(V - SV_S)dt. \tag{A2}$$

Substituting from (6) and (A1) reveals that the stochastic component of (A2) derives from only one source of risk—that arising from uncertainty over market shares, dz_θ. According to our assumptions, this risk is firm specific, or put differently, diversifiable, and therefore does not require a risk premium. Further, since the portfolio requires zero net initial investment, it must earn a zero expected return in equilibrium. Therefore, the deterministic part of (A2) must equal zero, which on substitution from (6) and (A1) yields (7)

References

Anderson, Ronald W. "The Determinants of the Volatility of Futures Prices." Working Paper no. 33. New York: Columbia Bus. School, Center for the Study of Futures Markets, June 1982.

Anderson, Ronald W., and Danthine, Jean-Pierre. "Hedger Diversity in Futures Markets. An Equilibrium Analysis of the Sources of Backwardation." Working Paper no. 71A. New York: Columbia Bus. School, Center for the Study of Futures Markets, June 1980.

———. "The Time Pattern of Hedging and the Volatility of Fu Working Paper no. 7. New York. Columbia Bus. School, C Study of Futures Markets, April 1981.

Black, Fischer. "The Pricing of Commodity Contracts." *J. Fe* (January/March 1976): 167–79.

Black, Fischer, and Scholes, Myron. "The Pricing of Options an Liabilities." *J.P.E.* 81 (May/June 1973): 637–54.

Bodie, Zvi, and Rosansky, Victor. "Risk and Return in Commodity Futures." *Financial Analysts J.* 36 (May/June 1980): 27–39.

Congressional Budget Office. "Agricultural Price Support Programs. A Layman's Guide." Washington: Government Printing Office, April 1976.

Constantinides, George M. "Market Risk Adjustment in Project Valuation." *J. Finance* 33 (May 1978): 603–16.

Cox, John C., and Ross, Stephen A. "The Valuation of Options for Alternative Stochastic Processes." *J. Financial Econ.* 3 (January/March 1976): 145–66.

Danthine, Jean-Pierre. "Information, Futures Prices, and Stabilizing Speculation." *J. Econ. Theory* 17 (February 1978): 79–98.

Feder, Gershon; Just, Richard E.; and Schmitz, Andrew. "Futures Markets and the Theory of the Firm under Price Uncertainty." *Q.J.E.* 94 (March 1980): 317–28.

Fischer, Stanley. "Call Option Pricing When the Exercise Price Is Uncertain, and the Valuation of Index Bonds." *J. Finance* 33 (March 1978): 169–76.

Holthausen, Duncan M. "Hedging and the Competitive Firm under Price Uncertainty." *A.E.R.* 69 (December 1979): 989–95.

Leland, Hayne E. "Theory of the Firm Facing Uncertain Demand." *A.E.R.* 62 (June 1972): 278–91.

Merton, Robert C. "Theory of Rational Option Pricing." *Bell J. Econ. and Management Sci.* 4 (Spring 1973): 141–83.

———. "Option Pricing When Underlying Stock Returns Are Discontinuous." *J. Financial Econ.* 3 (January/March 1976): 125–44.

———. "On the Pricing of Contingent Claims and the Modigliani-Miller Theorem." *J. Financial Econ.* 5 (November 1977): 241–49.

———. "On the Microeconomic Theory of Investment under Uncertainty." In *Handbook of Mathematical Economics*, edited by Kenneth J. Arrow and Michael D. Intriligator. Amsterdam: North-Holland, 1982.

Ross, Stephen A. "Return, Risk and Arbitrage." In *Studies in Risk and Return*, edited by Irwin Friend and James Bicksler. Cambridge, Mass.: Ballinger, 1975.

Rubinstein, Mark. "The Valuation of Uncertain Income Streams and the Pricing of Options." *Bell J. Econ.* 7 (Autumn 1976): 407–25.

Samuelson, Paul A. "Proof That Properly Anticipated Prices Fluctuate Randomly." *Indus. Management Rev.* 6 (Spring 1965): 41–49.

Sandmo, Agnar. "On the Theory of the Competitive Firm under Price Uncertainty." *A.E.R.* 61 (March 1971): 65–73.

Tomek, William C. "Empirical Analyses of the Demand for Food: A Review." Working Paper no. 77-8. Cornell Agricultural Econ. Staff Paper. Ithaca, N.Y.: Cornell Univ., April 1977.

Advertising as a Signal

Richard E. Kihlstrom and Michael H. Riordan

University of Pennsylvania

A great deal of advertising appears to convey no direct credible information about product qualities Nevertheless, such advertising may indirectly signal quality if there exist market mechanisms that produce a positive relationship between product quality and advertising expenditures. Two models of this phenomenon are presented. In each, advertising signals quality in the short run. The models differ in their treatment of the effect of advertising on long-run sales. In the first, all high-quality firms ultimately establish reputations for high quality whether they advertise or not. This is shown to imply that advertising can signal quality if and only if high-quality production requires investments in specialized assets that increase fixed costs but not marginal costs. In the second model, where nonadvertising firms never acquire a reputation for high quality, advertising might signal quality even if marginal production costs are somewhat lower for low quality. These conclusions closely
lel arguments previously made by Phillip Nelson.

I. Introduction

One observes a great deal of advertising that appears to
direct, credible information about product qualities. Phill
has argued in a series of papers (1970, 1974, 1978) that this sor
advertising can still be informative, but indirectly so, if there exist

---

Research support from the National Science Foundation, the Sloan Foundation, and the Center for the Study of Organizational Innovation is gratefully acknowledged Earlier versions of this paper were presented at the NBER/KGSM conference on Time and Uncertainty in Economics, held at Northwestern University in the spring of 1982 at the Economic Theory Workshop at the University of Pennsylvania, at Columbia University, at Princeton University, and at the Federal Trade Commission We are grateful for a number of helpful comments from participants at those seminars The paper has also benefited from comments by Jacques Cremer and an anonymous referee

[*Journal of Political Economy*, 1984, vol 92, no 3]

© 1984 by The University of Chicago All rights reserved. 0022-3808/84 9203-0009$01 50

market mechanisms that produce a positive relationship between product quality and advertising expenditures. In this case, consumers can correctly infer unobservable quality from observable advertising; that is, advertising signals quality.

Advertising can signal quality only if higher-quality firms find it profitable to invest in relatively more advertising than do lower-quality firms, and conversely. If advertising costs are the same for all brands, then the returns to advertising must be greater for higher quality. This feature distinguishes an advertising signal from the type of signal discussed by Spence (1973) for which asymmetry in the cost of the signal, not the return, is fundamental.

Nelson (1974) argues that the interaction of experience and repeat purchases can create an asymmetry in the returns to advertising capable of supporting advertising as a signal. He assumes that advertising expands the initial sales for all brands equally but increases repeat purchases of higher-quality brands relatively more. The differential impact of advertising on repeat purchases matters, because otherwise profits generated by advertising would be the same for all firms "unless costs behaved in a very peculiar manner" (1974, p. 753). Nelson's argument is intuitive and interesting. Unfortunately, he does not provide a complete model of market competition.

Richard Schmalensee (1978) has constructed a market model in which advertising is linked to repeat purchases through an ad hoc specification of consumer behavior. The model is loosely patterned on Nelson's ideas but does not really capture the informational role of advertising as a signal. In fact, Schmalensee exhibits equilibria in which low-quality firms profitably advertise more than high-quality firms. This possibility is certainly inconsistent with rational equilibrium behavior of consumers.

Here we develop two models of advertising as a signal that are more in the spirit of Nelson's argument. In both models, advertising is simply a conspicuous expenditure of resources by firms. Firms are price takers who by advertising influence the price at which their products can be sold. In the first model, we assume that consumers are uninformed about quality when making an initial purchase but fully informed about it when making subsequent purchases. Such situations will arise if communication between all consumers is perfect, so that each consumer's experience with initial purchases becomes common knowledge. With this information structure, advertising affects only the price at which initial sales are made, and it does so symmetrically for all firms. In this model, consumers interpret advertising as a signal of quality and rationally infer high quality when advertising expenditures are sufficiently large. When a firm signals by advertising, it demonstrates to consumers that its production cost and the demand for its product are such that advertising costs can be recovered. In order for advertising to be an effective signal, high-quality firms must be able to recover advertising costs while low-quality firms cannot. Consumers must also be well enough informed about costs to realize that advertising is profitable for high-quality but not for low-quality firms.

In this first model, repeat sales cannot be attributed to advertising. Because of the information structure, high-quality firms will make repeat sales even if they do not advertise. Thus, this model does not exactly capture the spirit of Nelson's argument. Nevertheless, advertising can arise in the model and the advertising is to some extent financed by the profits obtained from repeat sales, even though these sales are not due to advertising. As we show, however, advertising can be an effective signal in this model only if certain cost conditions are satisfied. As Nelson suggested, these cost conditions are a bit peculiar. They do, however, have an intuitive interpretation. Specifically, if high-quality production requires investment in specialized assets that increase fixed costs, but *not* marginal costs, relative to low quality, then there exist (possibly many) market equilibria in which advertising signals quality. If, however, higher quality raises marginal costs, then no such equilibria exist.

It is also shown that even without repeat purchases these same cost conditions alone can almost always generate an equilibrium in which advertising signals quality. In virtually all cases in which advertising equilibria exist with repeat purchases, they also exist without purchases. The one case in which advertising cannot ex repeat purchases is a borderline case, albeit a somewhat one, namely, the case in which marginal costs are the same and low quality. Except in this case, it is therefore not ne introduce the assumption of repeat purchases by consu learn from experience in order to explain advertising as a sig true, however (and this is shown in proposition 3 below), that the se of possible equilibria and the amount of advertising that can be financed in equilibrium will be affected if repeat sales become more important.

The final section of the paper briefly describes an alternative formulation of the model in which the informational assumptions are modified slightly. As noted above, the model discussed in the earlier sections of the paper assumes that high-quality firms that choose not to advertise ultimately establish reputations for high quality in spite of the fact that they do not advertise. They are, therefore, ultimately able to sell their output at the higher price obtainable in the high-quality market. In contrast, the model of the final section assumes that if all high-quality firms but one advertise, the nonadvertiser will never acquire a reputation for high quality and never make sales in the high-quality market. With this alternative formulation, advertising can be viewed as leading to repeat purchases. In this model, the returns to advertising by high-quality firms are higher than in the first model. As a consequence, in the model of the final section, advertising equilibria can exist even if marginal production costs are somewhat lower for low quality. Since the repeat purchases obtained by high-quality producers in the model of the final section are attributable to advertising, that model most adequately captures the spirit of Nelson's argument.

Klein and Leffler (1981) have also recently argued that advertising might signal quality under the assumption that low-quality firms enjoy both an average cost *and* a marginal cost advantage. Their model is most closely related to the model we describe in Section IV. The idea that fixed costs such as advertising might signal quality also appears in recent work by Wolinsky (1981, 1983). The primary focus of Wolinsky's work, however, is on prices as signals of quality in a model where consumers search for quality. Another model in which advertising signals quality is described in Johnsen (1976).[1] In Johnsen's work, some buyers are "experimenters" who will buy a product if a sufficiently large amount is spent advertising it. If these experimenters find the product to be of high quality, they recommend it to "imitators," who subsequently buy it. In this way, advertising by high-quality sellers leads to repeat sales but advertising by low-quality firms does not. For this reason, only high-quality firms advertise and advertising signals quality. Thus, Johnsen's model provides an interesting alternative to our formalization of Nelson's ideas.

The papers that make up the literature on prices as signals of quality are also obviously related to the present analysis. The most obvious connection is the informational asymmetry introduced by the fact that in these papers, as in ours, firms know quality while at least some consumers do not. Typically, these models assume that there is a subgroup of relatively well-informed buyers whose information is conveyed by the equilibrium price. Thus, the informational assumptions and the equilibrium concepts in this literature differ significantly from those investigated here. Contributions to this literature have been made by Farrell (1980), Chan and Leland (1982), and Cooper and Ross (1984).

II. Advertising Equilibrium with Repeat Purchases

Our model possesses a variety of special features:

1. There are only two qualities of the good available, high or low.
2. We consider advertising in a new market in which consumers

---

[1] We are grateful to Ed Prescott for this reference.

initially have no information about quality but eventually become perfectly informed through experience and reputation effects. We model these factors in a highly stylized way. The market evolves in two periods. In the first, consumers have no direct information on the qualities of competing brands but do observe advertising decisions. In the second, all consumers are perfectly informed about all previously offered brands. We are for now implicitly assuming that reputation and experience effects work equally quickly for high- and low-quality brands.

3. All entry takes place at the beginning of the first period. Firms make unalterable quality choices at that time. That is, a firm that produces high quality in the first period also produces high quality in the second period, and similarly for a low-quality firm. We explicitly rule out the possibility that a firm might initially produce high quality and then, having established a reputation, switch to low quality. This assumption makes sense if production requires investments in quality-specific assets that, once committed, have a low opportunity cost.

4. The acquisition of the technology required to produce high-quality goods requires a one-time expenditure of F_H in period 1. These expenditures together with advertising expenditures (if any) represent the fixed costs incurred by each high-quality producer Low-quality producers acquire their technology for an investment of F_L. Low-quality producers may also choose to advertise. Th costs equal F_L plus any amount they choose to spend on ad

5. The high-quality technology acquired by the expendit represented by the variable cost function $C(\cdot, H)$. Similarly, th ment F_L obtains a technology with the cost function $C(\cdot, L)$

We assume that storage is impossible, so that all the outpu period is produced in that period. Firms, therefore, make tw duction runs. Thus, a firm of quality q (where $q = H$ or $q = L$ producing x_1 units in period 1 and x_2 units in period 2 incurs costs of $C(x_1, q)$ in period 1 and $C(x_2, q)$ in period 2.

We assume that marginal costs are positive and increasing. Thus, $C_x > 0$ and $C_{xx} > 0$. Since C represents variable costs, $C(0, q) = 0$ for $q = H$ and $q = L$.

At this point we make no assumptions about the relative size of F_H and F_L or about the relationship between $C(\cdot, H)$ and $C(\cdot, L)$. We specifically do not assume that F_H exceeds F_L. Nor do we assume that variable costs are higher if the high-quality technology is used. Indeed, one of the primary objectives of our analysis is to investigate the nature of the relationship that can exist between variable and fixed costs of high- and low-quality production if advertising equilibria exist. We will, however, assume throughout that the minimum average cost of high-quality production is higher for high-quality producers. It is convenient for expositional reasons to postpone the statement of this assumption until more notation has been introduced and the formal definition of an advertising equilibrium has been presented.

6. In each period, there are two different markets: a high-quality market and a low-quality market. In the second period, firms are allocated to markets on the basis of true qualities. In the first period, firms are allocated to markets on the basis of advertising decisions. In order to enter the high-quality market, a firm must spend some minimum amount on advertising. Otherwise, a firm enters the low-quality market by default. Advertising is thus an "entry fee" into the high-quality market in the first period. As noted above, it becomes part of the firm's fixed cost.

7. Each market is competitive. Firms are price takers and prices clear markets. First-period entry continues until the total discounted profits from operations in both periods are zero in both markets. In order to simplify the analysis, we assume that there is a continuum of potential entrants. This assumption justifies the assumption of competition and eliminates the need to consider integer constraints on entry.

8. Buyers in the high-quality market are those who are willing to pay a premium, h, for the high-quality good. There are assumed to be m_H such buyers. The other m_L buyers are not willing to pay more for high quality.

The size of the premium, h, can be viewed as being determined endogenously and simultaneously with p_H and p_L, the high- and low-quality prices, respectively. In our exposition, we choose not to model the determination of h explicitly. However, one approach is to assume that for high-quality buyers the marginal rate of substitution between high and low quality is fixed at $k > 1$. Thus, the indifference curves between high- and low-quality consumption, x_H and x_L, respectively, are as shown in figure 1. Given p_L, the premium h will be given by $h = (k - 1)p_L$.

While we could proceed by making rather general assumptions about demand, we choose to simplify the discussion by assuming that in equilibrium each buyer purchases only one unit of whatever good (high or low quality) he chooses to buy.

9. We assume that all consumers and producers know the costs of producing high and low quality even though consumers do not initially know which firms use the high-quality technology. The consumers and producers also know how the high- and low-quality markets function.

Having discussed the basic features of the model, we can now begin to describe formally the signaling equilibrium. We first note that in

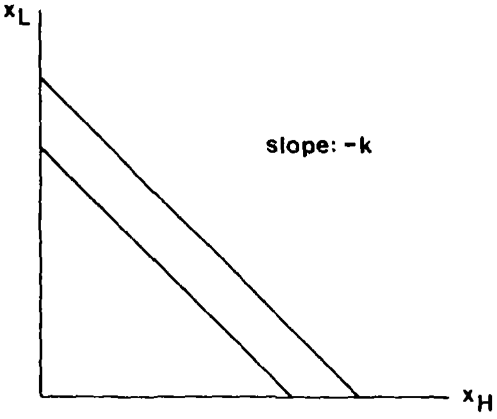


FIG. 1

each period t in which a firm produces a positive amount, the quantity supplied is that which maximizes period t profits gross of fixed costs. If the firm produces goods of quality q (where $q = H$ or $q = L$) and sells x_t units at a market price of p, its period t profits gross of fixed costs are $px_t - C(x_t, q)$. The maximizing x_t is denoted by $x(p, q)$ and satisfies

$$p = C_x(x(p, q), q). \tag{1}$$

We let

$$\phi(p, q) = px(p, q) - C(x(p, q), q)$$

be the maximized value of per period profits gross of fixed

Equilibrium is achieved in the low-quality market when the low-quality goods is such that discounted profits are zero satisfies

$$(1 + \delta)\phi(p_l, L) - F_l = 0,$$

where δ is the discount factor. We assume throughout that $\delta \geqq 0$ The case $\delta = 0$ can be identified with the case in which we drop the assumption of a second period during which consumers are informed. The role of this assumption in the analysis can therefore be studied by comparing the case $\delta = 0$ to the case of $\delta > 0$.

Similarly, the equilibrium price in the high-quality market is p_H, where

$$(1 + \delta)\phi(p_H, H) - (F_H + A) = 0, \tag{4}$$

and A denotes the amount of advertising necessary to enter the high-quality market.

We also have to impose a condition that will ensure that low-quality firms do not find it profitable to advertise and thereby enter the high-quality market in the first period. The profits obtained from such a strategy are $\phi(p_H, L) = \delta\phi(p_L, L) - (F_L + A)$, and these profits must be nonpositive; that is, we must have

$$\phi(p_H, L) + \delta\phi(p_L, L) - (F_L + A) \leq 0. \tag{5}$$

Finally, we must impose a condition that guarantees that the advertising expenditure, A, is not so large as to discourage high-quality firms from advertising in the first period. Specifically, the zero profits earned by high-quality firms must exceed the discounted profits they could earn by failing to advertise, selling at p_L in the first period and then selling at p_H in the second period. The profits earned from such a strategy are $\phi(p_L, H) + \delta\phi(p_H, H) - F_H$ and we must have

$$\phi(p_L, H) + \delta\phi(p_H, H) - F_H \leq 0. \tag{6}$$

An *advertising equilibrium* is a pair of prices (p_H, p_L) with $p_H > p_L$ and a positive advertising expenditure level A such that conditions (3)–(6) hold. We must of course also have $p_H - p_L < h$, where it will be recalled that h is the premium that m_H buyers are willing to pay for high quality. In such an equilibrium, each high-quality firm will produce $x(p_H, H)$ units and each low-quality firm will supply $x(p_L, L)$ units. There will then be $m_H/x(p_H, H)$ firms producing high-quality goods and $m_L/x(p_L, L)$ producers of low-quality goods.

Note that when conditions (3)–(6) hold, high-quality buyers who expect all advertisers to be high-quality producers will have rational expectations.

Before beginning the analysis of conditions under which advertising equilibria exist, we introduce some additional notation and state formally the hypothesis that the minimum average cost is higher in the high-quality industry. Specifically, we define p^0 by the equality

$$(1 + \delta)\phi(p^0, H) - F_H = 0. \tag{7}$$

Clearly, p^0 is the minimum average cost of production in the high-quality industry just as p_L is the minimum average cost of producing low quality. As such, p^0 is the equilibrium price that would prevail in the high-quality market if all consumers were informed in both periods and if $h + p_L \geq p^0$. Note that, since $A > 0$, p_H must exceed p^0.

We can now introduce the assumption that the minimum average cost of high-quality production exceeds that of low quality by simply assuming that $p^0 > p_L$.[2] This is a rather weak hypothesis. It does not imply that F_H exceeds F_L or that marginal costs are higher with the high-quality technology. It does, however, imply that either fixed costs or marginal costs are higher for high-quality production. (To be perfectly correct, if fixed costs are lower for high quality, p^o can exceed p_L only if marginal costs are sometimes higher for high quality.)

---

[2] Note that even if we did not assume p^0 exceeded p_L, this inequality would be implied by the conditions defining an advertising equilibrium. This is true since when p_H ex-

Observe that if p^o did not exceed p_L, we would find that all $m_H + m_L$ buyers, that is, both high- and low-quality buyers, would be supplied by high-quality producers. In such a case, the informational problems on which this paper focuses would fail to arise.

Note also that, when $\delta = 0$, $p_L < p^o$ implies that $\phi(p_L, H) - F_H < \phi(p^o, H) - F_H = 0$, so that (6) holds with a strict inequality.

We will now derive necessary conditions for the existence of an advertising equilibrium. As a first step, we observe that conditions (3) and (5) imply that A is such that $\phi(p_H, L) - \phi(p_L, L) \leqslant A$, and conditions (4) and (6) imply that $\phi(p_H, H) - \phi(p_L, H) \geqslant A$.

The first of these inequalities asserts that for low-quality producers the cost of advertising exceeds the increased profits due to advertising. The second inequality states that for high-quality producers the reverse is true. Of course, these two inequalities can hold simultaneously only if

$$\phi(p_H, L) - \phi(p_L, L) \leqslant \phi(p_H, H) - \phi(p_L, H). \tag{8}$$

When the inequality in (6) is strict, the inequality in (8) must also be strict. Inequality (8) plays a role in our analysis similar to that played in Spence's (1973) by the assumption that signaling costs decline as quality increases. In the present context, the signaling costs same for both high- and low-quality producers. But when the returns from signaling by advertising are at least as high quality producers as for low-quality producers. Inequality used to prove that no signaling equilibrium is possible if the cost of low-quality production is always lower than the mar of high-quality production. Formally, we have:

PROPOSITION 1: If $C_x(x, L) \leqslant C_x(x, H)$ for all x, and the inequality is strict for some interval of x values between $x(p_L, L)$ and $x(p^o, H)$, then no advertising equilibrium can exist. In addition, when $\delta = 0$, no equilibrium can exist when $C_x(x, L) = C_x(x, H)$, for all x.

PROOF: We will do the proof for the case in which $C_x(x, L) < C_x(x, H)$, for all x. A proof that applies to the more general case is easily constructed from this proof.

---

ceeds p^o (which, as noted above, it does in an advertising equilibrium), $p_L < p^o$ implies

$$\phi(p_L, H) + \delta\phi(p_H, H) - F_H > (1 + \delta)\phi(p^o, H) - F_H = 0,$$

in contradiction of (6).

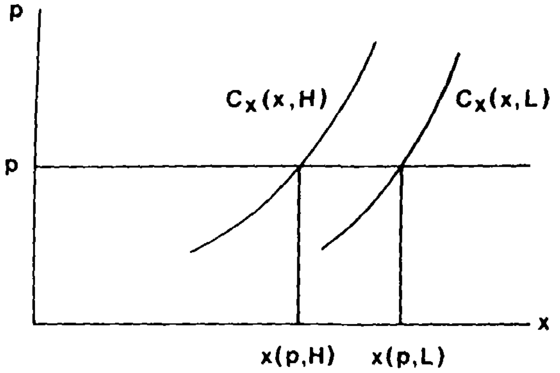


FIG. 2

First, note that for $q = H$ and $q = L$

$$\phi(p_H, q) - \phi(p_l, q) = \int_{p_l}^{p_H} \phi_p(p, q)dp. \qquad (9)$$

Also, for either q value, $\phi_p(p, q) = x(p, q)$. As figure 2 makes clear, if $C_x(x, L) < C_x(x, H)$, for all x, then

$$\phi_p(p, L) = x(p, L) > x(p, H) = \phi_p(p, H), \qquad (10)$$

for all p. Taken together, (9) and (10) imply that

$$\phi(p_H, L) - \phi(p_l, L) > \phi(p_H, H) - \phi(p_L, H); \qquad (11)$$

that is, (8) fails, and we cannot have an advertising equilibrium.

When $C_x(x, H) = C_x(x, L)$ for all x, the inequalities in (10) and (11) are replaced by equalities, so that (9) implies that (8) cannot hold as a strict inequality. If, in this case, $\delta = 0$, no advertising equilibrium could therefore exist. Q.E.D.

Inequality (11) is illustrated in figure 3, where $\phi(p_H, H) - \phi(p_l, H)$ is the cross-hatched area, B, and $\phi(p_H, L) - \phi(p_L, L)$ is the sum of the cross-hatched area, B, plus the shaded area, C.

One case in which $C_x(x, L) \leq C_x(x, H)$ occurs is where high-quality production requires the use of more expensive material or more highly skilled labor. Proposition 1 implies that in these cases advertising equilibria cannot exist.

Proposition 1 gives a sufficient condition for the nonexistence of advertising equilibria. It asserts that if an advertising equilibrium is to exist, the extra costs incurred in the production of high quality cannot result in a marginal cost curve for high-quality production that is always above the marginal cost curve for low-quality production. When $\delta = 0$—that is, when there is no second period in which all

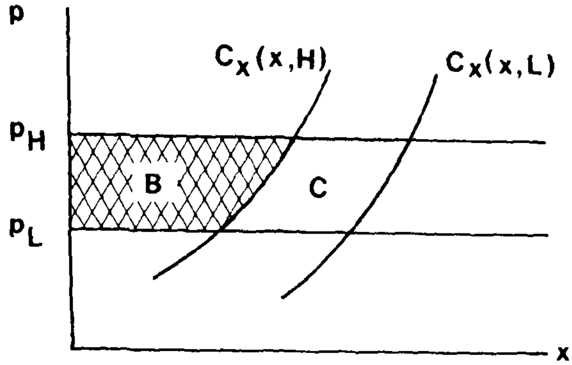


FIG. 3

consumers are informed—there is an additional case in which no advertising equilibrium can exist, namely, the case in which the marginal costs of producing high and low quality are always the same.

The argument used to obtain proposition 1 can be reinterpreted to conclude that an advertising equilibrium can exist only if, "on average," the marginal costs of low-quality production exceed or equal the marginal costs of high-quality production. In particular, (9) and the fact that $\phi_p(p, q) = x(p, q)$ can be used to interpret (8) as asserting that we must have $x(p, L) \leq x(p, H)$ more often than not. Because of figure 2, it is clear that $x(p, L) \leq x(p, H)$ is the predominant case if $C_x($ $C_x(x, H)$ is also predominant.

Having shown in proposition 1 that no advertising equili exist if the marginal cost of high-quality production is higher and never lower than that of low-quality produc now show that if low-quality producers have no marginal tage over high-quality producers, then advertising equilib ist. To do so, we first introduce some new notation.

In particular, we define $\underline{p}$ by

$$[(1 + \delta)\phi(\underline{p}, H) - F_H] - [\phi(\underline{p}, L) + \delta\phi(p_L, L) - F_L] = 0. \tag{12}$$

Note that when $\delta = 0$, (12) becomes

$$[\phi(\underline{p}, H) - F_H] - [\phi(\underline{p}, L) - F_L] = 0. \tag{12'}$$

When $\delta > 0$, we also define $\bar{p}$ by

$$[\phi(p_L, H) + \delta\phi(\bar{p}, H) - F_H] = 0. \tag{13}$$

For the present, we will simply assume that $\bar{p}$ and $\underline{p}$ exist and defer discussion of the conditions under which this is true. With the prices

$\bar{p}$ and $\underline{p}$ we associate the advertising levels $\bar{A}$ and $\underline{A}$ defined by $\bar{A} = (1 + \delta)\phi(\bar{p}, H) - F_H$, and $\underline{A} = (1 + \delta)\phi(\underline{p}, H) - F_H$, respectively.

Proposition 2, which follows, demonstrates that when low-quality production never has a marginal cost advantage, the numbers defined above can be given a simple interpretation. Specifically, if $p_L + h > \bar{p}$, then $\bar{A}$ will be the highest advertising level and $\bar{p}$ the corresponding highest high-quality price at which high-quality firms will be induced to undertake advertising in a free-entry equilibrium. This will be proved by showing that $\bar{p}$ is the maximum price satisfying (6), while $\bar{A}$ is the corresponding advertising level determined by the free-entry condition (4). Proposition 2 also demonstrates that $\underline{A}$ is the minimum advertising level, and $\underline{p}$ the corresponding minimum high-quality price, that dissuades low-quality firms from advertising in a free-entry equilibrium. This will be demonstrated by observing that when p_H is related to A by (1) condition (5) is equivalent to

$$[(1 + \delta)\phi(p_H, H) - F_H] - [\phi(p_H, L) + \delta\phi(p_L, L) - F_L] \geqslant 0, \quad (14)$$

and by noting that $\underline{p}$ is the minimum price satisfying (14).

Therefore, when low-quality production is not characterized by lower marginal cost than high-quality production, an advertising equilibrium must involve an advertising level A satisfying $\bar{A} \geqslant A \geqslant \underline{A}$ and a corresponding high-quality price, p_H, satisfying $\bar{p} \geqslant p_H \geqslant \underline{p}$. Clearly, $\bar{A} \geqslant \underline{A}$ and $\bar{p} \geqslant \underline{p}$ are necessary conditions for this to be possible. These inequalities are established in proposition 2. The proposition also establishes that these inequalities are sufficient for the existence of an advertising equilibrium as long as $\underline{p} < p_L + h$.

PROPOSITION 2: Assume that

$$C_x(x, L) \geqslant C_x(x, H) \quad (15)$$

for all x. If $\underline{p}$ exists, then $\underline{p} > p^0 > p_L$ and (14) holds at p_H if and only if $p_H \geqslant \underline{p}$. Also, when $\delta > 0$, $\bar{p}$ and $\underline{p}$ both exist, $\bar{p} \geqslant \underline{p}$ and $\bar{A} \geqslant \underline{A}$. If (15) holds for all x and the inequality is strict on some interval of x values between $x(p_L, L)$ and $x(\bar{p}, H)$, then when $\delta > 0$, $\bar{p} > \underline{p}$ and $\bar{A} > \underline{A}$. When (15) always holds with equality—that is, when $C_x(x, H) = C_x(x, L)$—then $\delta > 0$ implies $\bar{p} = \underline{p}$ and $\bar{A} = \underline{A}$.

Thus, if $\delta > 0$, if (15) holds for all x, and if $p_L + h > \underline{p}$, then for each $p_H \in [\underline{p}, \min(p_L + h, \bar{p})]$, (p_H, p_L, A) is an advertising equilibrium when A is defined by $A = (1 + \delta)\phi(p_H, H) - F_H$.

If, in addition, the inequality in (15) is strict on some interval of x values between $x(p_L, L)$ and $x(\bar{p}, H)$, then there are multiple advertising equilibria. If, however, $C_x(x, L) = C_x(x, H)$ for all x, then the advertising equilibrium is unique. If $\delta = 0$, and if there exists a p value such that $\phi(p, H) - \phi(p, L) \geqslant F_H - F_L$, then $\underline{p}$ exists. If in addition (15) holds with a strict inequality for all x and if $p_L + h > \underline{p}$.

then for each $p_H \in [\underline{p}, p_L + h]$, (p_H, p_L, A) is an advertising equilibrium when A is defined by $A = \phi(p_H, H) - F_H$.

PROOF: Let

$$\Psi(p) = [(1 + \delta)\phi(p, H) - F_H] - [\phi(p, L) + \delta\phi(p_L, L) - F_L].$$

Then $\Psi'(p) = (1 + \delta)x(p, H) - x(p, L)$. As noted in the proof of proposition 1, $x(p, H) \geq x(p, L)$ when $C_x(x, H) \leq C_x(x, L)$ for all x. Thus, $\Psi(p)$ is an increasing function of p, and $\Psi(p) \geq 0$ if and only if $p \geq \underline{p}$; that is, (14) holds at $p_H = p$ if and only if $p \geq \underline{p}$.

Since $p^o > p_L$, we have

$$\begin{aligned}\Psi(p^o) &= [(1 + \delta)\phi(p^o, H) - F_H] - [\phi(p^o, H) + \delta\phi(p_L, L) - F_L] \\ &< [(1 + \delta)\phi(p^o, H) - F_H] - [(1 + \delta)\phi(p_L, L) - F_L] \\ &= 0.\end{aligned}$$

Thus $\underline{p} > p^o > p_L$, if $\underline{p}$ exists.

We can now show that when $\delta > 0$, $\bar{p}$ exists and $\Psi(\bar{p}) \geq 0$. When combined with the continuity of Ψ and the fact just demonstrated that $\Psi(p^o) < 0$, $\Psi(\bar{p}) \geq 0$ implies that $\underline{p}$ exists. The inequality $\Psi(\bar{p}) \geq 0$ and the monotonicity of Ψ also imply that $\bar{p} \geq \underline{p}$.

To establish the existence of $\bar{p}$, we first observe that $p_L < p^o$ implies that $\phi(p_L, H) + \delta\phi(p^o, H) - F_H < 0$. We now use the easily established fact that under our hypothesis on $C(\cdot, H)$, $\phi(p, H)$ becomes infinite as p increases to infinity. Thus, there exists some p at which $\phi(p_L, H)$ $\delta\phi(p, H) - F_H > 0$. Since ϕ is continuous, these facts imp exists.

We begin the proof that $\Psi(\bar{p}) \geq 0$ by using the hypothesis $H) \leq C_x(x, L)$ for all x, to conclude (again using an argument the proof of proposition 1) that

$$\phi(\bar{p}, H) - \phi(p_L, H) \geq \phi(\bar{p}, L) - \phi(p_L, L).$$

Inequality (16) is equivalent to

$$\begin{aligned}&[(1 + \delta)\phi(\bar{p}, H) - F_H] - [\phi(p_L, H) + \delta\phi(\bar{p}, H) - F_H] \\ &\geq [\phi(\bar{p}, L) + \delta\phi(p_L, L) - F_L] - [(1 + \delta)\phi(p_L, L) - F_L]\end{aligned} \tag{17}$$

Using (3) and subtracting $[(1 + \delta)\phi(\bar{p}, H) - F_H]$ from both sides of (17), we obtain

$$\begin{aligned}-[\phi(p_L, H) + \delta\phi(\bar{p}, H) - F_H] &\geq [\phi(\bar{p}, L) + \delta\phi(p_L, L) - F_L] \\ &\quad - [(1 + \delta)\phi(\bar{p}, H) - F_H].\end{aligned} \tag{18}$$

Since $\bar{p}$ is defined by (13), the expression on the left side of inequality (18) is zero. The expression on the right is $-\Psi(\bar{p})$. Thus, (18) implies that $\Psi(\bar{p}) \geq 0$, so that $\bar{p}$ must exceed or equal $\underline{p}$.

The fact that $\bar{A} \geq \underline{A}$ now follows immediately from the fact that ϕ is an increasing function of p.

We have already observed that, given (4), (5) is equivalent to (14). Moreover, the monotonicity of $\Psi(p)$ implies that (14) is equivalent to $p_H \geq \underline{p}$. Finally, $p_H \leq \bar{p}$ is equivalent to (6) since $\phi_p > 0$. Therefore, we can conclude that when $\underline{p} < p_L + h$, (p_H, p_L, A) is an advertising equilibrium if $p_H \in [\underline{p}, \min(p_L + h, \bar{p})]$ and $A = \phi(p_H, H) - F_H$. With A so defined, $\bar{A} \geq A \geq \underline{A} \geq 0$.

When $\delta > 0$ and $p_H > \bar{p}$, (6) fails. We have already noted that when $p_H < \underline{p}$, the monotonicity of Ψ implies that (14) and hence (5) fails. As a result, p_H values below $\underline{p}$ or above $\bar{p}$ cannot be associated with advertising equilibria. If p_H exceeds $p_L + h$, the advertising equilibrium will fail to exist because there will be no buyers for the high-quality goods.

When $\delta > 0$ and $C_x(x, L) > C_x(x, H)$ for some interval of x values between $x(p_L, L)$ and $x(\bar{p}, H)$, inequalities (16)–(18) are strict, so that $\Psi(\bar{p}) > 0$ and $\bar{p} > \underline{p}$. Hence there are multiple equilibria.

When $C_x(x, L) = C_x(x, H)$ for all x, then (16)–(18) become equalities and $\Psi(\bar{p}) = 0$ so that $\bar{p} = \underline{p}$. In this case, the equilibrium must be unique.

If $\delta = 0$, then $\underline{p}$ can fail to exist, even if $C_x(x, L) > C_x(x, H)$ for all x. The problem is that the difference between $C_x(\cdot, L)$ and $C_x(\cdot, H)$ may be so small that there is no p at which the extra profits, $\phi(p, H) - \phi(p, L)$, earned by selling high quality rather than low, are sufficient to cover the additional fixed costs, $F_H - F_L$, of producing high quality rather than low. If, however, there exists p such that $\Psi(p) = [\phi(p, H) - \phi(p, L)] - [F_H - F_L] \geq 0$, then $\Psi(p^o) < 0$ and the continuity of Ψ implies that $\underline{p}$ exists. When $\underline{p}$ does exist, and $p_L + h > \underline{p}$, then $p_L < p^o$ implies (6), while (4) and $\Psi(p_H) \geq 0$ for all $p_H \geq \underline{p}$ imply (5). Thus, any p_H on the interval $[\underline{p}, p_L + h]$ can be associated with an advertising equilibrium in which $A = \phi(p_H, H) - F_H$. Q.E.D.

Proposition 2 asserts that there are multiple advertising equilibria when high-quality producers have a marginal cost advantage over low-quality producers. Each of these advertising equilibria are associated with a different level of advertising expense. This type of multiplicity is characteristic of signaling equilibrium models.

Note that, in the advertising equilibrium that arises when $A = \underline{A}$, (5) holds with an equality and all low-quality producers are indifferent about advertising. When, on the other hand, $A = \bar{A}$, (6) becomes an equality and high-quality producers are indifferent about advertising. In the case in which high- and low-quality producers face the same marginal cost curves, A must equal $\underline{A} = \bar{A}$, so that both (5) and (6) are equalities. In this "knife-edge" case, both high- and low-quality producers are indifferent about advertising.

As noted in proposition 1, this knife-edge case cannot give rise to

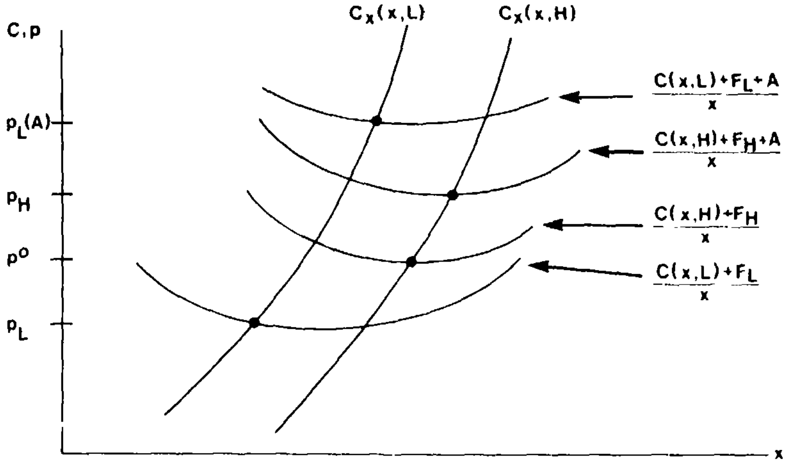


FIG. 4

an advertising equilibrium if $\delta = 0$, that is, if there is no second period in which all consumers are informed. Advertising equilibria are possible when $\delta = 0$ and when $C_x(x, H) < C_x(x, L)$ for all x. The required conditions are that $\underline{p}$ exists and $p_l + h > \underline{p}$. These conditions will hold if the marginal cost advantage of high quality is sufficiently great over the range of outputs $[x(p_l, L), x(p_l + h, L)]$. This ensures that $\Psi(p_L + h) > 0$ and hence that $\Psi(\underline{p}) = 0$ has a unique [illegible] with $\underline{p} < p_l + h$. Thus the hypothesis of a second period in [illegible] consumers are informed is not necessary for the existence [illegible] tising equilibria. High-quality producers must, however [illegible] sufficient marginal cost advantage over low-quality produc[illegible]

When $\delta = 0$ and $C_x(x, H) < C_x(x, L)$ for all x, an advertisin[illegible] rium can be described as in figure 4.

As indicated in figure 4, $p_l(A)$ and p_H are the minimum average cost levels associated with the low-quality and high-quality technologies, respectively, when A is included as a fixed cost. Note also that in figure 4 a minimum average cost reversal takes place. When advertising is included as a fixed cost, $p_L(A)$, the minimum average cost of producing low quality, exceeds p_H, the minimum average cost of producing high quality. This must be true because of the incentive compatibility condition (5) that guarantees that low-quality producers do not advertise. When, however, advertising is not included as a fixed cost, p_L, the minimum average cost of producing low quality, is lower than p^0, the minimum average cost of producing high quality. If $\delta = 0$, the strong form of inequality (8) can be interpreted as a necessary condition for the minimum average cost reversal. When

$C_x(x, L) \leq C_x(x, H)$ for all x, this necessary condition cannot hold and the reversal cannot occur. If, on the other hand, $C_x(x, L) > C_x(x, H)$ for all x, then the relatively high fixed cost, F_H, of high-quality production becomes progressively less important than marginal cost in the determination of the minimum average cost level as A increases. $\underline{A}$ is the advertising level and $\underline{p}$ is the p_H level at which reversal occurs; that is, at $\underline{A}$, $p_l(\underline{A}) = p_H = \underline{p}$. At A levels above $\underline{A}$, we have the situation depicted in figure 4; that is, $p_L(A) > p_H > \underline{p}$.

In the remainder of this section, we derive some comparative static results that describe the effect of the discount factor δ on $\underline{p}$, $\bar{p}$, $\underline{A}$, and $\bar{A}$. The discount factor can be interpreted as a measure of the importance of future sales made to informed buyers. It is shown that as future sales become more important—that is, as δ rises—$\underline{p}$ and $\bar{p}$ as well as $\underline{A}$ and $\bar{A}$ fall. This means that the amount of advertising and the price of high quality in the advertising equilibrium tend to decrease as future sales to informed buyers become more important. This is to be expected, since cases in which virtually no sales are made in the first period can be identified with very large discount rates δ.[3] In these cases, we would expect the price of high-quality goods to equal p'', the minimum average cost of high-quality production.

PROPOSITION 3: Assume that $C_x(x, L) > C_x(x, H)$ for all x. When $\delta > 0$, $\partial \bar{p}/\partial \delta < 0$, $\partial \bar{A}/\partial \delta < 0$, $\partial \underline{p}/\partial \delta < 0$, and $\partial \underline{A}/\partial \delta < 0$.

PROOF: Implicit differentiation of (12) and (13) yields

$$\frac{\partial \underline{p}}{\partial \delta} = - \frac{\phi(\underline{p}, H) - \phi(p_l, L)}{(1 + \delta)x(\underline{p}, H) - x(\underline{p}, L)}, \tag{19}$$

and

$$\frac{\partial \bar{p}}{\partial \delta} = - \frac{\phi(\bar{p}, H)}{\delta x(\bar{p}, H)}. \tag{20}$$

As has been shown earlier, the numerators and denominators of both of these expressions are positive when $C_x(x, H) < C_x(x, L)$ for all x. Since (4) implies that $\bar{A}$ and $\underline{A}$ rise with increases in $\bar{p}$ and $\underline{p}$, we also have $\partial \bar{A}/\partial \delta < 0$, and $\partial \underline{A}/\partial \delta < 0$. Q.E.D.

III. No-Advertising Equilibria

In an advertising equilibrium, prices are completely determined by cost conditions. When there is no advertising, however, supply and demand conditions interact to determine prices. In the no-advertising equilibrium, there are three markets in which supply must equal demand. In the second period, there are high- and low-quality markets as before. There are m_H buyers in the high-quality market and they are supplied by the high-quality producers. There are m_L buyers in the low-quality market and they are supplied by low-quality producers. In the first period, there is now a third market in which all producers sell at the same price to all $m_H + m_L$ buyers. We now define p_H^N to be the price in the high-quality market and p_L^N to be the price in the low-quality market. The price in the first-period market is denoted by p_α.[1]

---

[3] Mike Katz has pointed out to us that another effect is also at work here. Specifically, as δ rises, fixed costs per period fall. This also tends to lower $\underline{p}$ and $\bar{p}$.

In order to define a no-advertising equilibrium formally, we first note that the zero profit condition for low-quality firms is

$$\phi(p_\alpha, L) + \delta\phi(p_L^N, L) - F_L = 0, \tag{21}$$

and that the zero profit condition for high-quality firms is

$$\phi(p_\alpha, H) + \delta\phi(p_H^N, H) - F_H = 0. \tag{22}$$

When equilibrium is achieved in the low-quality market that meets in the second period, there will be n_L firms each supplying $x(p_L^N, L)$, and n_L will satisfy the supply equal demand condition

$$n_L x(p_L^N, L) = m_L. \tag{23}$$

Similarly, equilibrium in the high-quality second-period market requires that there be n_H firms, each producing $x(p_H^N, H)$ units, where
satisfies

$$n_H x(p_H^N, H) = m_H.$$

Finally, when supply equals demand in the first-period
and n_L must satisfy

$$n_H x(p_\alpha, H) + n_L x(p_\alpha, L) = m_H + m_L.$$

Using the fact that n_L and n_H satisfy (23) and (24), respectively,
can be rewritten as

$$m_H \frac{x(p_\alpha, H)}{x(p_H^N, H)} + m_L \frac{x(p_\alpha, L)}{x(p_L^N, L)} = m_H + m_L. \tag{25'}$$

Together, equations (21), (22), and (25') determine p_α, p_L^N, and p_H^N, that is, a no-advertising equilibrium is a vector of prices (p_α, p_L^N, p_H^N), for which the inequalities $p_L^N \leq p_H^N \leq p_L^N + h$ hold, and (21), (22), and (25') are satisfied.

---

[1] Allen (1983) provides a formulation of an equilibrium without advertising in which high-quality firms are rationed and prices serve as signals of quality

Note that if (p_α, p_L^N, p_H^N) satisfied (21), (22), and (25′) but p_L^N exceeded p_H^N, we would not be able to interpret the vector of prices as a no-advertising equilibrium. In this case, all high-quality producers would prefer to sell at p_l^N rather than p_H^N and total supply to the high-quality market would be zero, while supply in the low-quality market would be $n_H x(p_L^N, H) + n_L x(p_L^N, L)$. Observe that if $p^0 < p_L$, both high- and low-quality markets would be supplied by high-quality producers using the high-quality technology. Thus, we assume that p^0 exceeds p_L throughout the discussion of this section, as we did in the discussion of advertising equilibrium in the preceding section.

We now establish the following result:

PROPOSITION 4: For $h > 0$ sufficiently high, there exists a unique no-advertising equilibrium, (p_α, p_l^N, p_H^N), with

$$p_H^N > p_\alpha > p_l^N, \tag{26}$$

$$p_l^N < p_L < p_\alpha, \tag{27}$$

and

$$p_\alpha < p^0 < p_H^N < \bar{p}. \tag{28}$$

REMARK: Inequality (26) asserts that in a no-advertising equilibrium, the first-period price obtained by all sellers exceeds the price low-quality sellers will obtain in the second period but is lower than the second-period price high-quality sellers will obtain. For the purpose of interpreting inequality (27), it will be recalled that p_l is the price that would prevail in the low-quality market in an advertising equilibrium (p_l can also be interpreted as the minimum average cost of low-quality producers). It is therefore also the price that would prevail in the low-quality market if all buyers were informed in both periods. Inequality (27) thus asserts that p_l^N, the price received by low-quality producers in their second period of operation, is lower than their minimum average cost of production, p_L. They are compensated for this by the fact that p_α, the price they receive for their first-period output, exceeds the minimum cost of production p_l. To interpret inequality (28), we recall that p^0 is the minimum average cost of high-quality producers, and it would be the price prevailing in the high-quality market if there were no informational asymmetries. The reader will also recall that $\bar{p}$ is the maximum price possible in the high-quality market in an advertising equilibrium. The first inequality in (28) therefore asserts that the first-period price, p_α, received by high-quality producers is below their minimum cost of production, p^0. The second inequality asserts that the second-period price, p_H^N, received by these producers exceeds their minimum cost of production. The gains made by sales at p_H^N therefore compensate for first-period losses incurred because $p^0 > p_\alpha$. The final inequality in (28) guarantees that the price charged by high-quality producers in the second period of the no-advertising equilibrium is lower than the highest possible high-quality price in an advertising equilibrium. Since the price charged in any advertising equilibrium always exceeds p^o, (28) also implies that the first-period price obtained by high-quality producers is less than the price charged in all advertising equilibria.

Proof: Conditions (21) and (22) implicitly define p_L^N and p_H^N, respectively, as continuous (in fact, differentiable) functions of p_α. We denote these functions as

$$p_L^N = r_L(p_\alpha), \tag{21'}$$

$$p_H^N = r_H(p_\alpha). \tag{22'}$$

These can be substituted into (25′) to define the function

$$\Gamma(p_\alpha) = m_H \left\{ \frac{x(p_\alpha, H)}{x[r_H(p_\alpha), H]} - 1 \right\} + m_L \left\{ \frac{x(p_\alpha, L)}{x[r_L(p_\alpha), L]} - 1 \right\}.$$

Thus, given (21) and (22), (25′) is equivalent to $\Gamma(p_\alpha) = 0$.

Since

$$(1 + \delta)\phi(p^o, H) - F_H = \phi(p_L, H) + \delta\phi[r_H(p_L), H] - F_H = 0,$$

$$(1 + \delta)\phi(p_L, L) - F_L = \phi(p^o, L) + \delta\phi[r_L(p^o), L] - F_L = 0,$$

and $p^o > p_L$, we must have $r_H(p_L) > p^o > p_L$ and $r_L(p^o) < p_L$ · $r_H(p^o) = p^o$ and $r_L(p_L) = p_L$. These inequalities imply that

$$\Gamma(p_L) = m_H \left\{ \frac{x(p_L, H)}{x[r_H(p_L), H]} - 1 \right\} < 0,$$

and

$$\Gamma(p^o) = m_L \left\{ \frac{x(p^o, L)}{x[r_L(p^o), L]} - 1 \right\} > 0.$$

It is also easy to verify by implicit differentiation that $r_H'(p_\alpha) < 0$ and $r_L'(p_\alpha) < 0$ and to use these facts to show that $\Gamma'(p_\alpha)$ exists and is positive. There must, therefore, exist a unique p_α between p^o and p_L satisfying $\Gamma(p_\alpha) = 0$. This implies that (21), (22), and (25′) have a unique solution.

To establish that p_α cannot fall below p_L^N, we observe that if $p_\alpha < p_L^N < p_H^N$, we would have

$$\frac{x(p_\alpha, L)}{x(p_L^N, L)} < 1 \tag{29}$$

and

$$\frac{x(p_\alpha, H)}{x(p_H^N, H)} < 1, \tag{30}$$

so that the right side of (25′) would exceed the left side. Similarly, when p_α exceeds p_H^N, the inequalities (29) and (30) are reversed and the right side of (25′) is exceeded by the left side. Thus, p_α cannot exceed p_H^N. This argument proves (26).

To prove (27), we observe that (21) and (3) combine to imply that

$$\phi(p_\alpha, L) + \delta\phi(p_L^N, L) = (1 + \delta)\phi(p_L, L).$$

Since ϕ is an increasing function of p and since we have already established that $p_\alpha < p^o$, this equality could not hold unless (27) held.

To prove (28), we first observe that (22) and the definition of p^o combine to imply that

$$\phi(p_\alpha, H) + \delta\phi(p_H^N, H) = (1 + \delta)\phi(p^o, H).$$

Using once again the fact that ϕ increases with p and the fact already established that $p_\alpha < p^o$, we note that this equality can hold only if $p_\alpha < p^o < p_H^N$. The fact that $p_H^N < \bar{p}$ is a consequence of (13) and (22), which combine to yield

$$\phi(p_L, H) + \delta\phi(\bar{p}, H) = \phi(p_\alpha, H) + \delta\phi(p_H^N, H).$$

Since we already know that $p_\alpha > p_L$ and since ϕ is an increasing function, $\bar{p}$ must exceed p_H^N. Q.E.D.

IV. An Alternative Formulation

Proposition 1 of Section II asserts that an advertising equilibrium cannot exist if marginal production costs are higher for high-quality than for low-quality firms. In defining the advertising equilibrium, it was assumed that buyers are perfectly informed in the second period about the quality of goods supplied by all producers. Consequently, advertising in the first period does not generate any repeat purchases in the second.

In this section, we modify this informational assumption slightly to obtain an alternative formulation of Nelson's argument. We specifically assume that high-quality buyers communicate perfectly with other high-quality buyers but do not communicate at all with low-quality buyers. Consequently, if a high-quality firm makes a sale to a high-quality buyer in the first period, then all high-quality buyers learn of that firm's high-quality output. Furthermore, high-quality firms who sell to low-quality buyers are not discovered by high-quality buyers. Thus, advertising by high-quality firms does generate repeat purchases by high-quality buyers, but first-period sales to low-quality buyers result in future sales to low-quality buyers. As discussed in the introduction, this feature of the model makes it a more appropriate representation of Nelson's ideas.

As we shall see, this change in the informational assumption results in a higher return to advertising by high-quality firms. As a result, it will be possible to show that advertising equilibria can exist even if marginal production costs are lower for the low-quality technology. There will, however, be a limit to the marginal cost advantage that can be enjoyed by the low-quality firms. The restriction imposed by this limit will be reduced as δ rises. Thus, if repeat purchases are very important in the sense that δ is very large, then an advertising equilibrium can exist even if marginal production costs are significantly lower for low-quality producers.

We begin by describing the effect of the change in our informational hypothesis on the conditions defining equilibrium. In fact, the only condition affected is (6), which can now be dropped as a requirement for equilibrium.

It will be recalled that under the informational hypothesis of Section II, condition (6) guaranteed that high-quality firms advertised. Under that informational hypothesis, high-quality buyers knew that they could always sell at p_H in the second period even if they did not advertise. In the present section, that is no longer true. Consider the problem faced by a high-quality producer who must decide whether or not to advertise. If all other high-quality producers do advertise but he chooses not to, then virtually all high-quality buyers will be supplied by the other high-quality producers. As a result, the probability is zero that the nonadvertising high-quality producer makes a sale to a high-quality buyer. The probability is, therefore, one th
high-quality producer fails to establish a reputation for h
among high-quality buyers. Thus, this producer will sell
periods, and his profits will equal $(1 + \delta)\phi(p_L, H) - F_H$

$$(1 + \delta)\phi(p_L, H) - F_H < 0,$$

he will prefer to advertise. We now observe that condition
hold, since $p_L < p^o$ implies that

$$(1 + \delta)\phi(p_L, H) - F_H < (1 + \delta)\phi(p^o, H) - F_H = 0. \quad (31')$$

Thus, when $p_L < p^o$, and nonadvertisers receive p_L in both periods, an equilibrium is a pair of prices p_L and p_H and a positive advertising level A such that $p_H > p_L + h$ and (3)–(5) hold. We will refer to this new concept of advertising equilibrium as a Nelson advertising equilibrium. The mathematical requirements for a Nelson advertising equilibrium are less stringent than those inherent in an advertising equilibrium. Specifically, since (6) no longer holds, it is now no longer necessary that $\phi(p_H, H) - \phi(p_L, H) \geqslant A$. Thus, inequality (8) is not a necessary condition for the existence of a Nelson advertising equilibrium and proposition 1 no longer applies.

Inequality (31′) does, however, combine with (4) to imply

$$\left(\frac{1}{1+\delta}\right)A < \phi(p_H, H) - \phi(p_L, H). \tag{32}$$

Since (3) and (5) continue to hold, we must also have

$$\phi(p_H, L) - \phi(p_L, L) \leq A \tag{33}$$

as before. Using (32) and (33), we obtain

$$\phi(p_H, H) - \phi(p_L, H) > \left(\frac{1}{1+\delta}\right)[\phi(p_H, L) - \phi(p_L, L)], \tag{34}$$

a weaker version of (8), as a necessary condition for a Nelson advertising equilibrium. Since $1/(1 + \delta) < 1$, this inequality can hold even if (8) fails. Using (34) and graphical arguments that parallel those used to establish proposition 1, it is possible to prove an analogue of proposition 1 that asserts that there is a limit to how much $C_x(x, H)$ can exceed $C_x(x, L)$ if a Nelson advertising equilibrium is to exist.

It is also possible to obtain an analogue of proposition 2 that asserts that when $C_x(x, H)$ is not too much larger than $C_x(x, L)$, each p_H between $\underline{p}$ and $p_L + h$ can be identified with a Nelson advertising equilibrium in which $A = (1 + \delta)\phi(p_H, H) - F_H$. As before, each p_H above $\underline{p}$ will satisfy (14), which because of (4) is equivalent to (5). The need to restrict p_H to be less than $\bar{p}$ is eliminated because (6) is not a condition imposed on a Nelson advertising equilibrium. Thus, a Nelson advertising equilibrium will exist in all cases in which an advertising equilibrium exists. In fact, when $\delta > 0$, a Nelson advertising equilibrium will also exist in cases characterized by $C_x(x, H) \geq C_x(x, L)$, as long as $C_x(x, H) - C_x(x, L)$ is not "too large." A formal analysis of these cases (which is not presented here) reveals that the constraints imposed on $C_x(x, H) - C_x(x, L)$ become less stringent as δ increases, that is, as repeat sales become more important. For the extreme case of $\delta = 0$ in which there are no repeat sales, the two equilibrium concepts coincide.

V. Conclusions

Section II introduces an equilibrium concept in which apparently noninformative advertising signals quality. This equilibrium, which is called an advertising equilibrium, can be viewed as a formalization of the ideas of Nelson (1970, 1974, 1978). It is shown in proposition 1 of Section II that an advertising equilibrium cannot exist if marginal costs are uniformly lower in the low-quality industry. Repeat purchases, which are central to Nelson's argument, play a subsidiary role in an advertising equilibrium. Specifically, repeat purchases are necessary for the existence of an advertising equilibrium only when the marginal costs of high- and low-quality production are the same. Furthermore, although advertisers do make repeat sales, these sales cannot be said to be the result of advertising. For these reasons, the advertising equilibrium of Section II is a less than faithful representation of Nelson's ideas. Thus, Section IV introduces a variant of the advertising equilibrium. In this variant, which is called a Nelson advertising equilibrium, a slightly different informational hypothesis is employed to yield an equilibrium in which repeat purchases can be properly said to be caused by advertising. This feature makes it possible to obtain Nelson advertising equilibria even if marginal costs are lower in the low-quality industry.

We have chosen to analyze the advertising equilibrium of Section II in relatively complete detail. The analysis of the Nelson advertising equilibrium parallels the analysis of Section II. The parallels as well as the differences are described in Section IV. For purposes of comparison, a no-advertising equilibrium is described in Section III.

As in other signaling models, the advertising and Nelson advertising equilibria are, in general, nonunique. For example, in Section II it was shown that for each advertising level A in an interval $[\underline{A}, \bar{A}]$ there exists an advertising equilibrium in which A is the amount spent by each firm on advertising. With each A in $[\underline{A}, \bar{A}]$, there is a corresponding high-quality price p_H in the interval $[\underline{p}, \bar{p}]$. As the amount A required for entry to the high-quality market increases, p_H incre

Although we have not focused on the welfare prope
equilibria, we can show that it is possible to Pareto rank
ria. The exact argument depends, however, on the inte
advertising. On the one hand, advertising can be viewed
expenditure. In that case, the advertising equilibrium as
the lowest price for high quality and the lowest advert
Pareto dominant. As advertising and the high-quality pric
quality buyers are worse off because they pay a higher price for
goods they buy. On the other hand, advertising can be interpreted as a transfer from firms to consumers. Under this interpretation, advertising is not wasteful. When advertising increases, the higher price paid by high-quality buyers is used in part to pay for a transfer to other buyers. It is easy to show that the high-quality price rises by more than the amount required to pay for the advertising. In other words, the presence of advertising implies that the high-quality price must be above average production cost. Consequently, firms' production decisions are distorted away from average cost-minimizing levels, and the equilibrium high-quality price must also cover higher production costs. Average production costs are positively related to the amount of equilibrium advertising.

Formally, the increase in per firm revenues caused by the increase in A is $(\partial/\partial A)(xp) \triangleq (x + x_p p)(\partial p/\partial A)$, where $\partial p/\partial A$ is obtained by implicitly differentiating (4). Thus, $\partial p/\partial A = 1/\phi_p = 1/x$ and $(\partial/\partial A)(xp) = 1 + (x_p p/x)$, where $x_p p/x > 0$ is the elasticity of supply. As a result, $(\partial/\partial A)(xp) > 1$, and per firm revenues increase by more than the amount of the transfer A. So even if A is a transfer, the advertising equilibrium associated with $\underline{A}$ and $\underline{p}$ is Pareto superior to other advertising equilibria. Even this advertising equilibrium is Pareto inferior to the equilibrium obtained when high-quality buyers begin with perfect information. In this equilibrium, $A = 0$ and high-quality buyers pay p^o.

A welfare comparison of advertising equilibria with a no-advertising equilibrium is difficult and beyond the scope of this paper. However, there appear to be two competing effects. On the one hand, an advertising equilibrium results in a correct match of high-quality buyers with high-quality firms, while a no-advertising equilibrium does not. On the other hand, a no-advertising equilibrium has high-quality firms producing closer to minimum average cost. Which of these effects dominates appears to depend on how strongly high-quality buyers value high-quality goods.

References

Allen, Franklin. "Reputation, Product Quality and Equilibrium Rationing of Firms." Unpublished manuscript. Philadelphia: Univ. Pennsylvania, 1983.

Chan, Yuk-Shee, and Leland, Hayne E. "Prices and Qualities in Markets with Costly Information." *Rev. Econ. Studies* 49 (October 1982): 499–516.

Cooper, Russell, and Ross, Thomas W. "Prices, Product Qualities and Asymmetric Information: The Competitive Case." *Rev. Econ. Studies* (1984), in press.

Farrell, Joseph V. R. "Prices as Signals of Quality." Ph D. dissertation, Brasenose College, Oxford Univ., 1980.

Johnsen, Thore H. "*Advertising, Market Equilibrium and Information.*" Ph.D. dissertation, Carnegie-Mellon Univ., 1976.

Klein, Benjamin, and Leffler, Keith B. "The Role of Market Forces in Assuring Contractual Performance." *J.P.E.* 89 (August 1981): 615–41.

Nelson, Phillip. "Information and Consumer Behavior." *J.P.E.* 78 (March/April 1970): 311–29.

———. "Advertising as Information." *J.P.E.* 82 (July/August 1974): 729–54.

———. "Advertising as Information Once More." In *Issues in Advertising: The Economics of Persuasion,* edited by David G. Tuerck. Washington: American Enterprise Inst., 1978.

Schmalensee, Richard. "A Model of Advertising and Product Quality." *J.P.E.* 86 (June 1978): 485–503.

Spence, A. Michael. "Job Market Signaling." *Q.J.E.* 87 (August 1973): 355–74.

Wolinsky, Asher. "Prices as Signals of Product Quality." Economic Discussion Paper no. 232. Murray Hill, N.J.: Bell Laboratories, 1981.

———. "Prices as Signals of Product Quality." *Rev. Econ. Studies* 50 (October 1983): 647–58.

A Simple Theory of International Trade with Multinational Corporations

Elhanan Helpman

Tel-Aviv University and Massachusetts Institute of Technology

Using the idea that firm-specific assets associated with marketing, management, and product-specific R & D can be used to service production plants in countries other than the country in which these inputs are employed, I develop a simple general equilibrium model of international trade in which the location of plants in a differentiated product industry is a decision variable. The model is then used to derive predictions of trade patterns, volumes of trade, the share of intra-industry trade, and the share of intrafirm trade as functions of relative country size and differences in relative factor endowments.

I. Introduction

The role of multinational corporations in the conduct trade has grown over time and has reached very large pro the United States, for example, at the all-manufacturing national corporations accounted in 1970 for 62 percent of its ex ($22 billion out of $35 billion) and 34 percent of its imports ($10.5 billion out of $31 billion) (see U.S. Tariff Commission 1973, p. 322). It is therefore not surprising that the ramifications of their existence

---

I wish to thank Richard Caves for insightful discussions during my work on this project, as well as my wife Ruth for patiently listening to long lectures on the subject of this paper. Helpful comments on a previous version were provided by Eitan Berglas, Torsten Persson, Lars Svensson, and José Scheinkman. This is a revised and much simplified version of Harvard Institute of Economic Research Discussion Paper no. 961. The first version was written when I was a Visiting Professor in the Department of Economics at Harvard University.

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0001$01.50

are of major concern to international trade experts. Nevertheless, there exists no well-articulated theory that explains the conditions for their emergence or predicts under these conditions a structure of trade that comes close to observed trade patterns.

Existing general equilibrium theories of international trade have been developed without explicit treatment of the multinational corporation. The discussion of direct foreign investment in Caves (1971) (and the work that followed from it) is an exception. There are many treatments of the multinational corporation in a partial equilibrium framework (see Caves 1982, chap. 2), but they shed only limited light on a central problem of trade theory, namely, the explanation of trade patterns. We are in need of a theory that describes conditions under which firms find it desirable to shift activities to foreign locations and that is able to predict the pattern of trade that emerges under these conditions. Foundations of such a theory are proposed in this paper, with the following important features: (*a*) there are differentiated products, economies of scale, and monopolistic competition; and (*b*) there exist inputs (e.g., management, marketing, and R & D) that can serve product lines without being located in their plants.

In this paper the theory deals with single product firms. I deal with horizontal and vertical integration in Helpman (1983*b*). Firms maximize profits and make, therefore, cost-minimizing location choices of product lines. This feature brings about the emergence of multinational corporations as a result of the tendency of factor rewards to differ across countries. Here the emphasis is on one source of pressure on relative factor rewards—differences in relative factor endowments. Transport costs and tariffs are assumed away, so that production facilities are not established in order to save transport costs or in order to produce behind tariff walls. Other reasons for multinationality, such as tax advantages of various forms, are also not considered.

Apart from describing in a general equilibrium system conditions under which firms choose to become multinational, the theory provides an explanation of trade patterns in which the multinational corporations play a central role. There is intersectoral, intra-industry, and intrafirm trade. The last trade component has become of major importance in recent years (see U.S. Tariff Commission 1973, chap. 3; Buckley and Pearce 1979).

In order to bring out as clearly as possible the value added of the theory, simplifying assumptions are used throughout. The next section provides a description of the basic model. The structure of an equilibrium in an integrated world economy is described in Section III. Then, in Section IV, the features of the integrated world equilibrium are used in order to describe the relationship between factor endowments and trade patterns. The behavior of the volume of trade is analyzed in Section V and the behavior of the shares of intra-industry and intrafirm trade is analyzed in Section VI. The last section is devoted to concluding remarks.

II. The Basic Model

For the purpose of the current study I employ a two-sector modified version of the now standard model of international trade in differentiated products. Preferences are assumed to be identical everywhere and representable by a homothetic utility function $u(Y, U_x)$, where Y is the consumption level of a homogeneous product and $U_x = u_x(\cdot)$ is the subutility level attained in the consumption of differentiated products. The function $u_x(\cdot)$ depends on the specification of preferences for a differentiated product; they can be, for example, of the Dixit and Stiglitz (1977) type or of the Lancaster (1979) type (see Helpman [1983*a*, sec. 8] for a description). In both cases a demand function facing a producer of a single variety can be derived; in the Dixit-Stiglitz case this demand function is of the constant elasticity *type (assumed to be larger than one), while in the Lancaster case its* elasticity depends on commodity prices and the number of varieties available to consumers (and this elasticity is always larger than one).

It is assumed that there are two factors of production: labor, L, and a general purpose input, H, whose special role in the production of differentiated products will be explained below. The homogeneous product is produced by means of a standard linear ho production function with the associated unit cost functi where w_i is the reward to factor i. A producer of the h product has to employ all inputs in the same location. In equilibrium the price of the homogeneous product, tak numeraire, equals unit costs:

$$1 = c_Y(w_L, w_H).$$

The structure of production of differentiated products is more complicated. A firm that wants to produce a given variety has to hire the general purpose input H and adapt it at a cost in order to make it suitable for the production of this variety. Once adapted, the input becomes a firm-specific asset in the sense used by Williamson (1981), and it is tied to the entrepreneurial unit. However, this firm-specific input can serve many plants and it need not be located within a plant in order to serve its product line. In particular, it can serve plants that are located in different countries (see Hirsch [1976] for a similar assumption). Inputs that fit this description are management, distribution, and product-specific R & D. The importance of this type of asset in the operation of multinational corporations is described in

Caves (1982, chap. 1). Clearly, in practice, combinations of inputs are required in order to generate such assets; here this aspect is simplified by assuming that only H can serve this purpose.

Let $l(x, h_X)$ be the quantity of labor required to produce x units of a variety of the differentiated product in a single plant when h_X units of H have been adapted for its particular use. A possible form for this function is $l = f_p + g_1(x, h_X)$, where $f_p > 0$ and $g_1(\cdot)$ is positively linear homogeneous. Here f_p generates a plant-specific fixed cost and the variable cost component exhibits constant returns to scale. More generally, I assume that $l(\cdot)$ is the inverse of an increasing-returns-to-scale production function in which h_X is essential for production. Let also $g(w_L, w_H, h_X)$ be the minimum costs required in order to adapt h_X to the desired variety, where $g(\cdot)$ is associated with a nondecreasing-returns-to-scale production function. Then the firm's single plant cost function is

$$C_X(w_L, w_H, x) = \min_{h_X} [w_L l(x, h_X) + g(w_L, w_H, h_X) + w_H h_X].$$

This function obviously has the standard properties of cost functions associated with increasing-returns-to-scale production functions. One can also define cost functions for larger numbers of plants. The point worth noting, however, is that the firm or corporation has fixed costs that are corporation specific but not plant specific (they consist of hiring h_X and adapting it), it has plant-specific fixed costs, and it has plant-specific variable costs. The assumption that $l(\cdot)$ is the inverse of an increasing-returns-to-scale production function implies that it pays to concentrate production in a single plant unless there are transportation costs or differences across location in product prices. Since impediments to trade are not considered in this paper, the single plant cost function described above is relevant for what follows. All varieties have the same cost structure.

It is assumed that there is Chamberlinian-type monopolistic competition in the differentiated product sector. Hence, as is well known, in this case firms equate marginal revenue to marginal costs and free entry brings about zero profits in every firm. In a symmetrical equilibrium these two conditions can be written as:

$$px = C_X(w_L, w_H, x) \tag{2}$$

and

$$R(p, n) = \theta(w_L, w_H, x), \tag{3}$$

where p is the price of every variety of the differentiated product; $R(\cdot)$ is average revenue divided by marginal revenue, and it measures the degree of monopoly power (it is a constant under the Dixit-Stiglitz specification of preferences); n is the number of varieties available to consumers; and $\theta(\cdot)$ is average costs divided by marginal costs, using $C_X(\cdot)$, and it measures the degree of returns to scale in the production of differentiated products (see Helpman 1981).

The formal conditions of industry equilibrium (1)–(3) that were described above are identical to the conditions used in existing models of trade in differentiated products (see Helpman 1983*a*). The important difference lies in the interpretation of the technology available to corporations in the differentiated product industry. As in most trade theory I will assume that factors of production do not move across national borders. However, because of the technology available in the differentiated product industry, the firm-specific asset h_X can serve product lines in plants that are located in countries other than the country in which h_X is located, and the specificity of h_X implies that arm's-length trade in its services is an inferior organizational form to an integrated firm (see Klein, Crawford, and Alchian 1978). This is precisely the feature that brings about the emergence of multinational corporations. We will call the country in which h_X and the entrepreneurial center are located the parent country of the corporation and the country in which the subsidiary is located the host country.

III. Equilibrium in an Integrated World Economy

As a first step toward the study of international trade be
mies of the type described in the previous sections, I de
section the symmetrical equilibrium of an integrated wo
The features of the integrated world economy will the
identify patterns of cross-country distributions of the wor
ment of labor and the *H* factor, which generate certain tr
and volumes of trade. This particular link provides valua
tion because differences in factor endowments can be asso
differences in relative country size and differences in rela
endowments, two variables that play a major role in empirical stud
We will study trade patterns and volumes of trade for a fixed-size
world economy.

In a symmetrical equilibrium of an integrated world economy factor prices are the same everywhere, and all the corporations that operate in the sector that produces differentiated products have a similar structure; every corporation produces one variety, but there is no overlap in varieties produced by two different corporations. they employ the same quantity of the *H* factor and the same quantity of labor; they charge the same price for every variety and produce the same quantity of each one of them. Free entry into the industry brings profits down to zero. The number of corporations n is treated as a continuous variable. This is a reasonable approximation when n is a large number.

Apart from (1)–(3) the equilibrium conditions consist of equilibrium conditions in factor markets and in commodity markets. The equilibrium conditions in commodity markets depend on the specification of preferences, and we will not present them because no use is made of them in what follows (for an example see Helpman 1981). It is only important to remember that the upper tier utility function $u(Y, U_X)$ is homothetic. The equilibrium conditions in factor markets are:

$$a_{LY}(w_L, w_H)y + A_{LX}(w_L, w_H, x)n = L, \tag{4}$$

$$a_{HY}(w_L, w_H)y + A_{HX}(w_L, w_H, x)n = H, \tag{5}$$

where $a_{iY}(w_L, w_H) = \partial c_Y(w_L, w_H)/\partial w_i$, $i = L, H$, is the cost-minimizing input of factor i per unit output of the homogeneous product; $A_{iX}(w_L, w_H, x) = \partial C_X(w_L, w_H, x)/\partial w_i$, $i = L, H$, is the cost-minimizing input of factor i in a representative corporation in the differentiated product industry, and L and H are the total quantities of labor and the H factor available. The quantity $A_{HX}(\cdot)$ consists of h_X plus any other quantity of H that might be required as an input in the process that converts h_X into a firm-specific asset. Condition (4) assures equilibrium in the labor market, and (5) assures equilibrium in the H market. Conditions (1)–(5) plus an equilibrium condition in commodity markets (e.g., that the demand for Y equal its supply) determine the equilibrium values of factor rewards (w_L and w_H), the price of differentiated products (p), the output level of a single variety of the differentiated product (x), the output level of the homogeneous good (y), and the number of corporations in the differentiated product industry (n), which equals the number of varieties available to consumers.

For what follows I make the natural assumption that in this equilibrium the homogeneous product is labor intensive relative to the differentiated product; that is,

$$\frac{a_{LY}}{a_{HY}} > \frac{A_{LX}}{A_{HX}}.$$

Under this assumption the equilibrium distribution of employment across sectors can be described by means of figure 1. The vector $\mathbf{O\bar{E}}$ represents the endowment of factors of production, the vector **OQ** represents employment in the differentiated product industry, while OQ' represents employment in the homogeneous good industry. The line BB' represents an equal factor cost line; its slope equals relative factor rewards. It is tangent to an isoquant of the homogeneous good

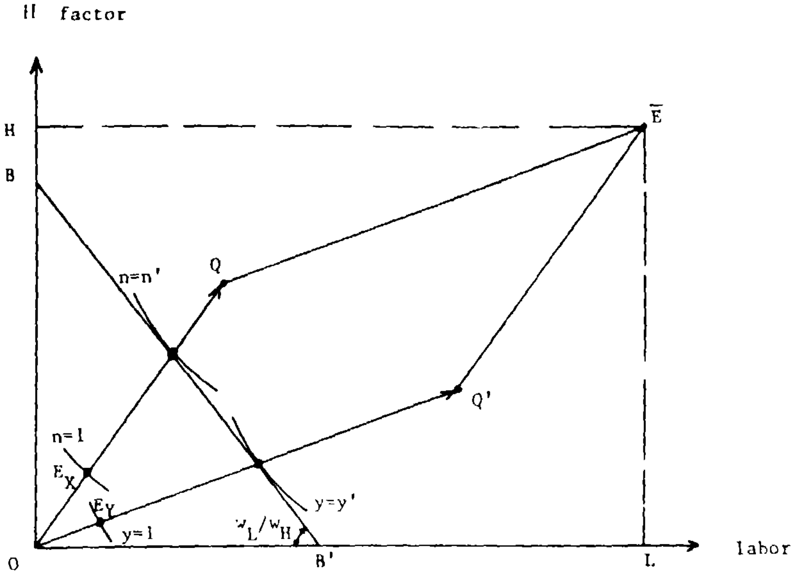


FIG. 1

at its intersection point with OQ'. At the intersection of BB' with OQ the equal factor cost line is also tangent to an isoquant, but one of different nature. This isoquant can be recovered from th tion by means of the set $\{(L_X, H_X) \mid \exists(w_L, w_H) > 0$ such th $[A_{LX}(w_L, w_H, x), A_{HX}(w_L, w_H, x)]\}$. This is the collection makes possible the production level of a single corporati be represented by a regularly shaped isoquant, labeled n 1. Now draw an entire family of isoquants by a radial e contraction of this single corporation isoquant. Every isoq family represents a different number of corporations, equally on a ray from the origin. Thus, at the intersection of BB' with OQ there is an isoquant belonging to this family that is labeled $n = n'$ Because of its definition this isoquant is tangent to BB'.

Finally, observe that using the single corporation isoquant we can calculate the number of corporations that can operate in equilibrium with inputs represented by a point on OQ by dividing the distance of the point from the origin by $\overline{OE}_X$. In a similar way the output of the homogeneous good that is obtained in equilibrium by an input combination represented by a point on OQ' can be calculated by dividing the distance of the point from the origin by $\overline{OE}_Y$. This completes the description of the integrated world equilibrium that is necessary for what follows.

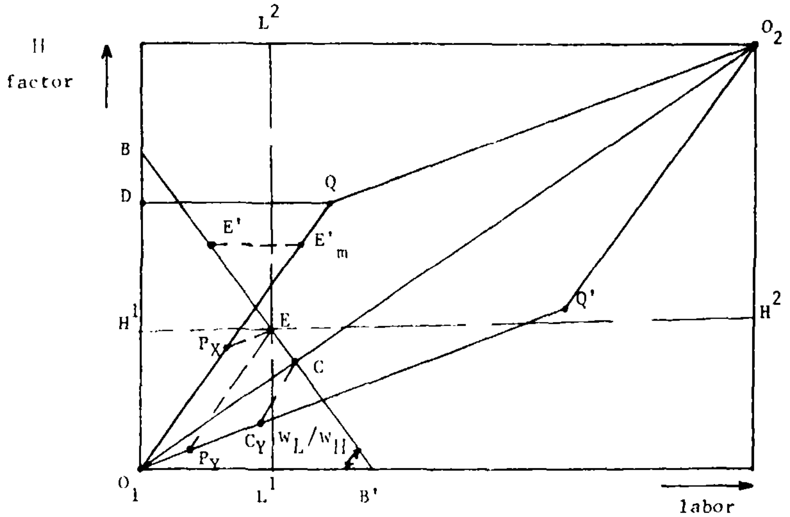


FIG. 2

IV. The Pattern of Trade

In the standard Heckscher-Ohlin two-country, two-good, two-factor model in which there are no factor intensity reversals and preferences are homothetic and identical across countries, the set of endowment allocations can be divided into two subsets. In the interior of one subset there is factor price equalization and no specialization in production, and every country exports the good whose production makes relatively intensive use of the factor with which the country is relatively well endowed. In the interior of the other subset every country pays a lower reward to the factor of production with which it is relatively well endowed and a higher reward to the other factor of production, at least one country specializes in the production of the good which is a relatively heavy user of its cheaper factor of production, and the pattern of trade is the same as in the former subset. If figure 2 (which is a box-diagram reproduction of fig. 1 with O_j representing the origin for country j) were to describe feasible allocations across countries for a standard Heckscher-Ohlin type economy, then the set with factor price equalization would be represented by O_1QO_2Q' and the other set by its complement.[1]

---

[1] See Dixit and Norman (1980, chap. 4), who also deal with a case in which there is factor intensity reversal.

The pattern of trade that emerges in the present model is much richer than described above. It is useful to describe it by starting with intercountry factor allocations in the set O_1QO_2Q' of figure 2. Because of the symmetry in structure it is sufficient to analyze endowment points above the diagonal O_1O_2; this way country 1 is the H-factor-rich country while country 2 is the labor-rich country. Allocations in this set were analyzed in Helpman (1981).

Take, for example, the factor endowment point E in figure 2. This point describes an allocation (L^1, H^1) of labor and the H factor to country 1 and an allocation (L^2, H^2) of labor and the H factor to country 2. It is straightforward to see that with this world structure there is an equilibrium with factor price equalization whose cross-country aggregation looks the same as the integrated world equilibrium described in the previous section. In this equilibrium corporations based in one country have no incentive to open subsidiaries in the other country in order to locate product lines there. Assuming that under these circumstances all operations of a corporation are concentrated in the parent country, the output level of the homogeneous product in country j and the number of corporations in the differentiated product industry in that country can be solved from the following factor market equilibrium conditions:

$$a_{LY}y^j + A_{LX}n^j = L^j, \tag{6}$$

$$a_{HY}y^j + A_{HX}n^j = H^j. \tag{7}$$

Here the input-output coefficients in the homogeneous p
tor and labor and H-factor use per corporation in the di
product industry are taken from the equilibrium of th
world economy, because factor prices, product prices, and
variety are the same in both cases. This is indeed an equili
solution of (y^j, n^j) from (6)–(7) is nonnegative for $j = 1, 2$.
of course, the case for every point in O_1QO_2, as is demonstra
broken-line parallelogram drawn from the particular E in figure 2

Now observe that because profits are zero all income is factor income. Hence, by drawing through E a line BB' whose slope is w_L/w_H, we show the cross-country income distribution. Relative incomes can be read off as follows: Let C be the intersection point of BB' with the diagonal O_1O_2. Then the relative income of country 1 is $\overline{O_1C}$ divided by $\overline{CO_2}$. In fact, by a proper choice of units, $\overline{O_1C}$ represents the income level of country 1 and $\overline{CO_2}$ represents the income level of country 2. Since both countries have the same spending pattern, country 1 consumes a proportion s^1 of the world's output y, where s^1 is its share in world income. Hence with a line through C parallel to O_1Q its consumption of the homogeneous product can be represented by $\overline{O_1C_Y}$.

where C_Y is the intersection point of this line with O_1Q'. Since production y^1 is represented by $\overline{O_1P_Y}$, country 1 imports the homogeneous product. Finally, since trade is balanced, this means that country 1 is a net exporter of differentiated products.

I have shown that in the set of factor allocations O_1QO_2 the intersectoral pattern of trade is the same as in the Heckscher-Ohlin model. Here, however, there is also intra-industry trade in differentiated products. Country j produces n^j varieties of the differentiated product and it exports them to its trading partner. Hence, the pattern of trade that emerges is the same as in the models of trade in differentiated products that were developed in recent years. To summarize, for factor endowments in the set O_1QO_2 free trade generates no incentive for the formation of multinational corporations. The structure of trade is the same as in recent models of trade in differentiated products; the intersectoral trade pattern is explained by differences in relative factor endowments while intra-industry trade is explained by monopolistic competition in differentiated products.

The theory proposed in this paper takes on interest because it can identify and analyze the implications of circumstances in which corporations find it profitable to establish subsidiaries abroad. This theory associates multinational corporations with the ability of firms to exploit cross-country differences in factor prices by shifting activities to the cheapest locations. Generally speaking, this theory can be applied to differences in factor prices that result from many different sources However, in what follows it is applied to potential differences in factor rewards that arise from differences in relative factor endowments.

It is clear from conventional theory and from the previous discussion that factor endowment points above O_1QO_2 lead to unequal factor prices *if* firms have to employ all factor inputs in the same country. Suppose that under these circumstances the H factor is cheaper in country 1 and labor is cheaper in country 2. Now consider what happens when a corporation need not employ all labor and H at a single location; for simplicity also assume that no labor is used in the process that adapts h_X to the particular variety produced by the corporation. Clearly, under these circumstances corporations wish to choose country 1 as their parent country and they wish to open subsidiaries in country 2. These desires reduce the demand for labor in country 1 and increase it in country 2, and they increase the demand for the H factor in country 1 and reduce it in country 2. An equilibrium is attained when either factor prices are equalized or country 1 becomes the parent country of all corporations (with unequal factor prices all H producing differentiated products are located in the H-cheap country). When factor price equalization obtains, there are many equilibrium configurations with various degrees of foreign involvement of the corporations in the differentiated product industry, just as there are many configurations in the factor price equalization set O_1QO_2. In the latter case, factor price equalization is achieved without invoking the possibility that corporations can decentralize their activities geographically. In the case under current examination factor price equalization can be achieved *because* companies can decentralize their activities geographically. There are many ways in which the decentralization can be made consistent with equilibrium. The rule to be adopted below is to consider equilibria in which foreign labor employment is as small as possible, which amounts to considering equilibria with the smallest number of multinational corporations.[2]

Start by considering factor allocations that are in the set O_1DQ of figure 2. I argue that endowment points in this set lead to equilibria with factor price equalization and the emergence of multinational corporations (I maintain in the figure the assumption that labor is not used in the adaptation process). Clearly, for endowment points in this set there are no equilibria in which factor prices are equalized, and every firm employs its factors of production in a single location. Hence, multinational corporations have to emerge. The only question that remains, therefore, is whether their emergence brings about factor price equalization. Take, for example, the endowment point E' in figure 2. If all the resources of country 1 are employed in the production of differentiated products and its corporations employ in
foreign country the amount of labor $\overline{E'E'_m}$, where E'_m is th
tion point with O_1Q of a horizontal line drawn through
aggregate world equilibrium corresponds to the equilib
integrated world economy. In this discussion E' is the
point and E'_m is the employment point. The existence of
corporations enables the employment point to differ from
ment point. The distance $\overline{O_1E'_m}$ represents the number of
tions that are based in country 1 (n^1) and the distance $\overline{E'_mQ}$ represents the number of corporations that are based in country 2 (n^2) (the total number is the same as in the integrated world equilibrium) More precisely, since $y^1 = 0$ and $y^2 = y$, the number of corporations that are based in country j, n^j, and the employment of labor in country 2 by subsidiaries of country 1–based multinationals, L^f, are obtained from the following factor market clearing conditions:

$$A_{LX}n^1 = L^1 + L^f, \quad a_{LY}y + A_{LX}n^2 = L^2 - L^f. \tag{8}$$

$$A_{HX}n^1 = H^1, \quad a_{HY}y + A_{HX}n^2 = H^2. \tag{9}$$

---

[2] This choice can be justified as a long-run equilibrium of a dynamic adjustment process in which it is costly to shift plants abroad.

However, the number of varieties produced in country j does not equal n^j; the number of varieties produced in country 1 is smaller than n^1 and the number of varieties produced in country 2 is larger than n^2. The precise difference depends on the size of the labor force L^f employed by subsidiaries. In fact the number of varieties produced in country j, M^j, $j = 1, 2$, is

$$M^1 = n^1 - \frac{L^f}{A_{LX}}, \quad M^2 = n^2 + \frac{L^f}{A_{LX}}.$$

To summarize, we have seen that endowment points in the set O_1DQ lead to an equilibrium with factor price equalization and the emergence of multinational corporations. Under my assumption about locational tendencies of corporations, in this set country 1 specializes in the production of differentiated products and it serves as a base for the multinational corporations. Country 1 imports the homogeneous product, and there is intra-industry trade in differentiated products. Part of the intra-industry trade is carried out by multinationals. It is also easy to see that the set O_1DQ can be divided into two subsets such that in one subset country 1 is a *net exporter* of differentiated products and in the other subset country 1 is a *net importer* of differentiated products. Finally, there exists intrafirm trade whose nature is discussed below.

The existence of intrafirm trade is, of course, well documented in the empirical literature (see, e.g., U.S. Tariff Commission 1973, chap. 2; Buckley and Pearce 1979). This takes the form of imports of the parent firm from its subsidiaries as well as exports of the parent firm to its subsidiaries. Much of this trade stems from vertical integration with which one cannot deal satisfactorily by means of the framework employed in this study, but which can be dealt with in a proper extension (see Helpman 1983*b*). However, there is one genuine component of intrafirm trade that is well represented by this model, namely, the invisible exports of the parent to its subsidiaries of services of the H input. Observe that because of the zero-profit condition (2) labor costs are lower than the revenue obtained from sales. This means that the multinational corporation is making "profits" in its subsidiary, because the subsidiary hires labor only in the host country. This means that the profits of the subsidiary are just sufficient to cover the costs of the H input, which is hired in the parent country. The difference between revenue and labor costs of all subsidiaries is $\mu px - w_L L^f$, where $\mu = L^f/A_{LX}$ is the number of multinational corporations. This can be considered to be either profits repatriated by the parent firms or payments by the subsidiaries for services rendered by the parent firms. From an economic point of view the second interpretation is the appropriate one. Hence, $\mu px - w_L L^f$ represents intrafirm trade.

In order to proceed with the analysis of trade patterns figure 3

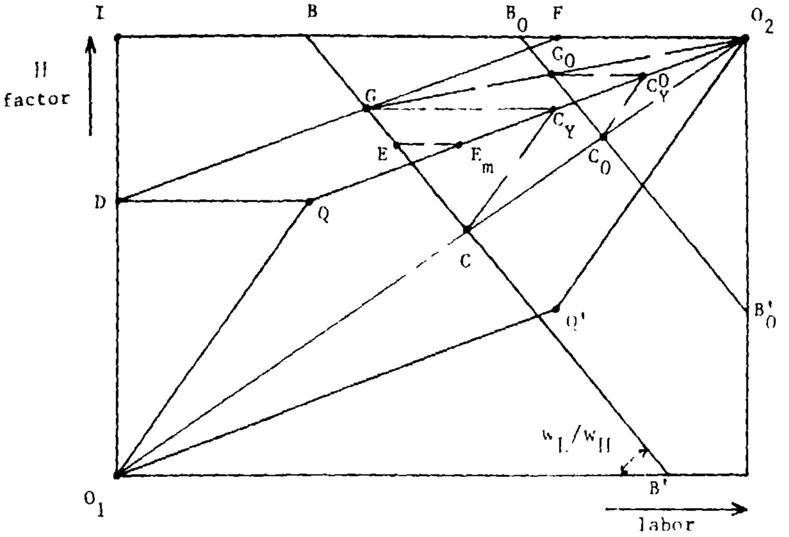


FIG. 3

reproduces the basic features of figure 2 and contains a further division of the set above O_1QO_2. Now consider endowment points in the parallelogram O_2QDF. Following the discussion of endowment points in the set O_1DQ it is clear that endowment points in O [illegible] equilibria with factor price equalization and multina[illegible] tions that are homomorphic to the integrated world ec[illegible] rium. Thus, if E is the endowment point, then E_m is a te[illegible] ment point when the techniques of production of th[illegible] world equilibrium are used, and $\overline{EE_m}$ represents labor e[illegible] country 1–based multinationals in country 2. It is also [illegible] figure 3 that endowment points in the set DIF can have no equ[illegible] with factor price equalization, for in order to use the integrated world equilibrium production techniques country 1 has to employ in country 2 a quantity of labor larger than total employment required in the differentiated product industry. This cannot take place, because only in this sector can corporations shift activities to other countries. Hence, for endowment points in the set DIF, all corporations that produce differentiated products are based in country 1 and all product lines are located in country 2.[3] Country 1 exports services of the H input and imports all the differentiated products from country 2. It may import or export the homogeneous product.

---

[3] This is necessarily so when production is homothetic. It seems that the reverse pattern of specialization is possible in the absence of homotheticity.

Returning to endowment points in the set O_2QDF, observe that there are two possible patterns of trade depending on the location of the endowment point. For the endowment point E and the employment point E_m, country 1 imports the homogeneous product as well as varieties of the differentiated product that are produced by its subsidiaries, and it exports the other varieties as well as services of the H input. The only part of this assertion that requires elaboration is that country 1 imports the homogeneous product. This can be seen as follows. Let BB' be the equal cost line that passes through E. Then its intersection point with the diagonal O_1O_2, that is, point C, represents each country's income level. Draw through C a line parallel to O_1Q and let C_Y be its intersection point with O_2Q. Then $\overline{O_2C_Y}$ represents the consumption level of the homogeneous product in country 2 while $\overline{O_2E_m}$ represents the output level of the homogeneous product in country 2. Hence, country 2 exports the homogeneous product and country 1 imports it. It is clear from figure 3 that the same pattern of trade emerges for all endowment points on BB' that belong to O_2QDF.

Now consider endowment points in O_2QDF that lie on B_0B_0'. For the distribution of income represented by this line the consumption of the homogeneous product is represented by $\overline{O_2C_Y^0}$ (constructed in the same way as above). It is clear, however, that only at endowment points below G_0 does country 2 export the homogeneous product, while at endowment points above G_0 country 1 exports the homogeneous product. At point G_0 (as well as at point G) there is no trade in the homogeneous product. Generally, at endowment points that are in O_2QDF but below GO_2, the pattern of trade is as the one described for point E. On the other hand, at endowment points that belong to O_2QDF but are above GO_2, country 1 exports the homogeneous product as well as services of the H input and some varieties, while country 2 exports only varieties that are produced by subsidiaries of country 1–based multinationals. A further division into subsets in which country 1 is a net exporter or a net importer of differentiated products is also possible.

We have identified five sets of endowment distributions relevant for the study of trade patterns (O_1QO_2, O_1DQ, O_2QDG, O_2GF, and DIF). In four of them the equilibrium requires the existence of multinational corporations. Overall, they represent a rich collection of trade patterns with features that seem to fit reality better than those provided by existing trade theories.

V. The Volume of Trade

I have described in the previous section possible trade patterns and how they are related to the distribution of the world's endowment of

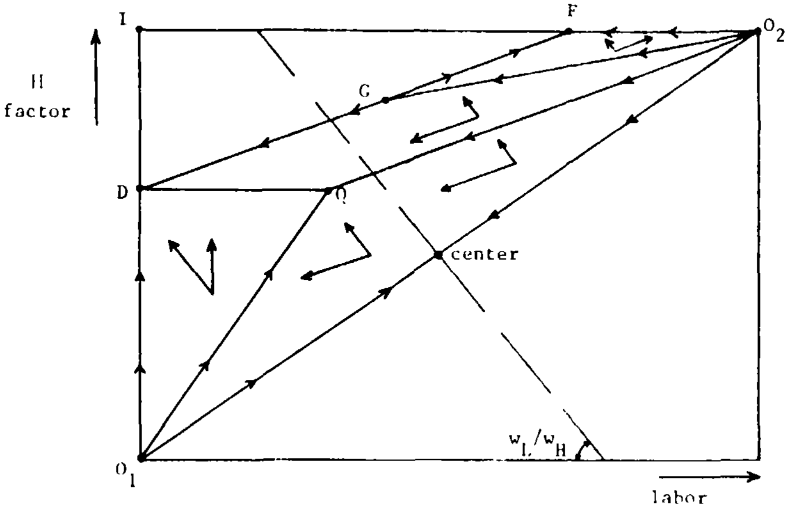


FIG. 4

factors of production. As is clear from the partition of the endowment set into subsets in which certain patterns of trade obtain (see fig. 3), the pattern of trade depends on two factors: (*a*) relative country size in terms of GNP and (*b*) the difference in relative factor endowments. For example, if country 1 is relatively small and has a relatively high endowment of the *H* factor, then the endowment p
O_1DQ, there exist multinational corporations that are
1, country 1 exports differentiated products and m
to its subsidiaries, and it imports differentiated prod
mogeneous good. Differences in relative factor endo
relative sizes of the countries are observable econo
major interest. I provide, therefore, in this section a d
effects that these two variables have on the volume of
findings are summarized by figure 4 in which the arrows
directions in which the volume of trade increases.

Start by considering endowment points in the set O_1QO_2. In this region there are no multinational corporations. The volume of trade is defined in the usual way as the sum of exports, where the summation is over countries and sectors. Because of balanced trade, this is equal in a two-country world to twice the exports of one of the countries. In this set country 1 exports only differentiated products, so that the volume of trade is:

$$V = 2pxs^2M^1 \quad \text{for } E \in O_1QO_2, \qquad (10a)$$

where M^1 is the number of varieties produced in country 1 and s^2 is the share of country 2 in world income. Hence, for fixed relative country size the volume of trade increases with the number of varieties produced in the H-factor-rich country, and for a fixed number of varieties produced in country 1 the volume of trade increases with the relative size of country 2. Since M^1 is fixed on lines parallel to O_2Q, the arrows inside the set O_1QO_2 and on O_2Q in figure 4 describe directions in which the volume of trade increases.

On the diagonal O_1O_2 we have $M^1 = s^1M$, where $M = n$ is the number of varieties produced in the world economy. Upon substitution into (10a) this yields:

$$V = 2pxMs^1s^2 \quad \text{for } E \in O_1O_2, \tag{10b}$$

implying that the volume of trade is largest when the countries are of equal size and is smaller the more unequal the size of countries. This is also described by the arrows in figure 4.

On the line O_1Q we have $pxM^1 = s^1Z$, where Z is world income and spending and $M^1 = n^1$. Hence, upon substitution into (10a):

$$V = 2Zs^1s^2 \quad \text{for } E \in O_1Q \tag{10c}$$

and the volume of trade rises the more equitable is the world's distribution of income. In figure 4 country 1 is of smaller size than country 2 at endowment points on O_1Q (which is, of course, not always the case); therefore the volume of trade increases in the northeastern direction as indicated by the arrow. The behavior of the volume of trade in the set O_1QO_2 described above is a generalization of propositions 5 and 6 in Helpman (1981).

For endowment points in O_1DQ country 2 exports the homogeneous good and M^2 varieties of the differentiated product. Some of the varieties it exports are produced by country 2–based firms while μ of them are produced by subsidiaries of country 1–based multinational corporations. The same trade pattern exists in O_2QDG, except that in this set all the varieties exported by country 2 are produced by country 1–based multinationals (i.e., $M^2 = \mu$). This means that the volume of trade can be represented as follows:

$$V = 2(y^2 - Y^2 + pxs^1M^2) \quad \text{for } E \in (O_1DQ) \cup (O_2QDG). \tag{11}$$

For endowment points in O_1DQ the homogeneous product is produced only in country 2, so that $y^2 = y$. Also, since $Y^2 = s^2y = (1 - s^1)y$, upon substitution into (11) we obtain:

$$V = 2s^1(y + pxM^2) \quad \text{for } E \in O_1DQ. \tag{12}$$

This means that when the relative country size is given the volume of trade increases with the number of varieties produced in country 2, whereas when the number of varieties produced in country 2 is given, the volume of trade increases with the relative size of country 1, as shown in figure 4.

At endowment points in the set O_2QDG all the varieties exported by country 2 are produced by subsidiaries of country 1–based multinationals; that is, $M^2 = \mu$. Also $Y^2 = s^2 y$ and $s^2 Z = y^2 + w_l \mu A_{lX}$, where $Z = pxn + y$. Combining these with (11), we obtain:

$$V = 2[(1 - s^1)pxn + px\mu(s^1 - \theta_{lX})] \quad \text{for } E \in O_2QDG, \qquad (13a)$$

where θ_{lX} is the share of labor costs in the production of differentiated products. Hence, when the relative country size is given the volume of trade increases with the widening of the difference in relative factor endowments (with μ) if and only if the share of country 1 in world income exceeds the share of labor costs in the differentiated product industry. The arrows in O_2QDG that indicate the increase in the volume of trade in the northwestern direction (see fig. 4) are drawn on the assumption $s^1 > \theta_{lX}$ (which always holds for endowment points close to O_2). It is also clear from (13a) that given μ the volume of trade decreases as the relative size of country 1 increases. However, μ is constant on lines parallel to O_2Q. Therefore, the volume of trade increases in the southwestern direction as indicated by the arrows drawn in region O_2QDG and on DG. Finally, on O_2G we have $M^2 = \mu$, $y^2 = Y^2 = s^2 y$, and $s^2 Z = y^2 + w_l \mu A_{lX}$, which yield upon substitution into (11):

$$V = \frac{2s^1 s^2 px(Z - y)}{w_l A_{lX}} \quad \text{for } E \in O_2G$$

implying that the volume of trade declines the mo
tries are in relative size. This is indicated by the arrow

It remains to consider endowments in O_2GF. In th
2 exports only differentiated products produced b
country 1–based multinationals, and the volume of trade

$$V = 2pxs^1\mu \quad \text{for } E \in O_2GF, \qquad (14)$$

implying that for given μ the volume of trade rises with the relative size of country 1 and that for given relative country size it rises with μ. On O_2F μ is proportional to s^2, which makes the volume of trade larger the closer the endowment is to F (because on O_2F, $s^1 > s^2$). These features are also represented by arrows in figure 4.

In summary, figure 4 represents a fairly detailed description of the relationship between factor endowments and the volume of trade. It shows that in some sense the larger the difference in relative factor endowments the larger is the volume of trade. On the other hand, relative country size has an ambiguous effect on the volume of trade.

VI. Intra-Industry and Intrafirm Trade

In this section I investigate the dependence of the shares of intra-industry and intrafirm trade on cross-country differences in relative factor endowments. In all cases, except for region O_1QO_2, the investigation is restricted to fixed relative sizes of the two countries.

The volume of intra-industry trade is defined as the total volume of trade minus the sum over all sectors of the absolute value of the difference between imports and exports. In the current model this reduces to

$$V_{I-I} = 2px \min (s^1M^2, s^2M^1). \tag{15}$$

The definition of the volume of intrafirm trade is more complicated. Exports of the parent firms of services of the H factor are undoubtedly part of this volume of trade. The problem arises with the treatment of the finished differentiated products. If parent firms serve as importers of the finished products manufactured by their subsidiaries, then this appears in the data as intrafirm trade, and similarly if subsidiaries serve as importers of the differentiated products manufactured by parent firms. In some cases the treatment of these flows of goods as intrafirm trade has no economic justification because it is more the consequence of bookkeeping practices than a true economic calculus. In the present model there is no natural choice—much depends on the implicit assumptions about the marketing technology. I choose, therefore, to define intrafirm trade as trade in the services of the H factor. Hence,

$$V_{I-F} = px\mu - w_LL^f = a\mu, \tag{16}$$

where $a = w_HA_{HX} > 0$.

For endowment points in O_1QO_2 we have $s^1M^2 < s^2M^1$, so that using (10a) and (15) we calculate the share of intra-industry trade to be

$$S_{I-I} = \frac{s^1M^2}{s^2M^1} = \frac{s^1n^2}{s^2n^1} \quad \text{for } E \in O_1QO_2. \tag{17}$$

This is shown in Helpman (1981, proposition 4) to be a declining function of the difference in relative factor endowments. The share of intrafirm trade is zero at endowment points that belong to O_1QO_2 because in this region there are no multinational corporations.

Using (12), we obtain:

$$S_{I-I} = \frac{px \min (s^1M^2, s^2M^1)}{s^1(y + pxM^2)} \quad \text{for } E \in O_1DQ. \tag{18}$$

Since one can show that $M^1/M^2 > s^1/s^2$ on O_1Q, then for a given relative country size the share of intra-industry trade rises with the difference in relative factor endowments as we start moving the endowment allocation along an equal income line from a point on O_1Q in the northwestern direction. If country 1 is small enough, however, this share reaches a maximum and declines with further redistributions that increase the gap in relative factor endowments. This means that with the emergence of multinational corporations the share of intra-industry trade may be positively or negatively related to differences in relative factor endowments. Using (16), the share of intrafirm trade in O_1DQ can be represented by

$$S_{i-f} = \frac{a(\mu/2)}{s^2pxM^1 + a\mu} \quad \text{for } E \in O_1DQ, \qquad (19)$$

where $s^2pxM^1 + a\mu$ is the volume of exports from country 1. Therefore, for given relative country size a widening of the difference in the H to L ratio between country 1 and country 2 increases the share of intrafirm trade (because it increases μ and reduces M^1).

For an endowment point in O_2QDG we have $M^2 = \mu$, so that (15), (16), and (13a) imply:

$$S_{i-i} = \frac{px \min(s^1\mu, s^2M^1)}{(1 - s^1)pxM + px\mu(s^1 - \theta_{LX})} \quad \text{for } E \in O_2QDG, \qquad (20)$$

$$S_{i-f} = \frac{a(\mu/2)}{(1 - s^1)pxM + px\mu(s^1 - \theta_{LX})} \quad \text{for } E \in O_2QDG \qquad (21)$$

Assume $s^1 > \theta_{LX}$ in this region. Then, for a given relative
starting from O_2Q the share of intra-industry tr
creases in the difference between the H to L ratio
country 2, and it may reach a maximum and decline
pattern is similar to that in O_1DQ. The share of intra
other hand, is larger the larger the difference in re
dowments (given relative country size).

In region O_2GF country 2 exports only differentiated
Therefore, because of balanced trade, at endowment points that belong to it $s^1\mu > s^2M^1$. Using this relationship as well as (11), (15), and (16), we find that on lines in O_2GF that represent constant relative country size, the share of intra-industry trade declines with increasing differences in relative factor endowments and the share of intrafirm trade is constant.

The broad picture that emerges from this analysis is that for given relative country size the share of intrafirm trade is larger the larger the difference in relative factor endowments, but that in the presence of multinational corporations no clear-cut relationship exists between the share of intra-industry trade and differences in relative factor endowments.

VII. Concluding Comments

I have developed in this paper a general equilibrium theory of international trade in which multinational corporations play an essential role. This theory can identify and analyze the implications of circumstances in which corporations find it profitable to become multinational. These corporations are well-defined economic entities, they possess firm-specific assets, they engage in monopolistic competition, and they play an active role in foreign trade. The theory explains the simultaneous existence of intersectoral trade, intra-industry trade, and intrafirm trade. Despite the relative richness of the theory it needs further extensions and elaborations in order to deal with the wide range of problems that are at the heart of international economics. An extension of the current theory to horizontally as well as vertically integrated corporations is presented in Helpman (1983*b*). This extension generates more realistic patterns of resource allocation without altering the fundamental properties of trade patterns that were derived in the current study. In particular, integrated multinational corporations end up having production facilities in parent as well as in host countries, and the existence of vertical integration brings about intrafirm trade both in H services and in intermediate inputs. This realism is achieved at a substantial cost in terms of the complexity of the theory of the firm. Nevertheless, I believe the benefit-cost ratio to be larger than one. The current theory can also explain cross-country penetration of multinational corporations as a result of impediments to trade (such as transport costs or tariffs). This is evident from the fact that the establishment of a new plant for the same variety requires additional fixed costs but saves the costs associated with trade impediments and does not require the hiring of new H factors. Hence, for sufficiently high impediments, cross-country penetration is expected.

References

Buckley, Peter J., and Pearce, Robert D. "Overseas Production and Exporting by the World's Largest Enterprises: A Study in Sourcing Policy." *J. Internat. Bus. Studies* 10 (Spring/Summer 1979): 9–20.

Caves, Richard E. "International Corporations: The Industrial Economics of Foreign Investment." *Economica* 38 (February 1971): 1–27.

———. *Multinational Enterprise and Economic Analysis*. Cambridge: Cambridge Univ. Press, 1982.

Dixit, Avinash K., and Norman, Victor. *Theory of International Trade*. Cambridge: Cambridge Univ. Press, 1980.

Dixit, Avinash K., and Stiglitz, Joseph E. "Monopolistic Competition and Optimum Product Diversity." *A.E.R.* 67 (September 1977): 297–308.

Helpman, Elhanan. "International Trade in the Presence of Product Differentiation, Economies of Scale and Monopolistic Competition: A Chamberlin-Heckscher-Ohlin Approach." *J. Internat. Econ.* 11 (August 1981): 305–40.

———. "Increasing Returns, Imperfect Markets, and Trade Theory." In *Handbook of International Economics*, vol. 1, edited by Ronald W. Jones and Peter B. Kenen. Amsterdam: North-Holland, 1983. (*a*)

———. "Multinational Corporations and Trade Structure." Mimeographed. Cambridge, Mass.: M.I.T., December 1983. (*b*)

Hirsch, Seev. "An International Trade and Investment Theory of the Firm." *Oxford Econ. Papers* 28 (July 1976): 258–70.

Klein, Benjamin; Crawford, Robert G.; and Alchian, Armen A. "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process." *J. Law and Econ.* 21 (October 1978): 297–326.

Lancaster, Kelvin. *Variety, Equity, and Efficiency* New York. Columbia Univ. Press, 1979.

U.S. Tariff Commission. *Implications of Multinational Firms for World Trade and Investment and for U.S. Trade and Labor.* Washington: Government Printing Office, 1973.

Williamson, Oliver E "The Modern Corporation. Origins, Evolution, Attributes." *J. Econ. Literature* 19 (December 1981)· 1537–68

Restrictions on Price Advertising

Michael Peters

University of Toronto

In this paper the effects of advertising restrictions in an industry producing an undifferentiated product are analyzed. When the product is undifferentiated, advertising conveys information about price. The standard intuition suggests that such prohibitions will raise buyer search costs, making demand less elastic. This raises price and profits in an industry. Advertising restrictions provide an indirect means of coordinating pricing strategies in this framework. Here we consider the cost side of the same argument. The efficiency of firms in an industry is allowed to vary and information about costs is private. In this environment efficient firms are hurt by advertising restrictions, since lower advertised prices constitute their most effective competitive weapon. We show that when the number of buyers and sellers is finite, there is a *nonzero* probability that advertising restrictions will reduce every firm's price. This occurs when all firms are more efficient than their competitors expect them to be. Further, when the number of buyers and sellers is infinite, it is possible that average price charged in an industry can fall after a ban on advertising

I. Introduction

When the primary function of advertising is to convey information, advertising restrictions are thought to have a relatively simple effect. Without advertising, it becomes more costly for some buyers to find low-priced suppliers. This reduces the demand elasticity faced by firms and raises equilibrium prices and profit rates. There is empirical evidence in support of this argument. For example, Benham (1972) shows that in states where advertising of eyeglasses was prohibited, average prices were more than double what they were in states where such restrictions did not exist. In addition, Boyer (1974) shows that

I would like to thank Ken Chan, Curt Eaton, Barbara Spencer, and two anonymous referees, one of whom suggested the analysis of Sec III

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0001$01.50

profit rates are negatively correlated with advertising in industries where product differentiation is difficult.

This argument is based on an analysis of demand behavior. In this paper I wish to look at the supply side of the same argument. Supply considerations may reverse the previous reasoning. The argument is reasonably simple. In industries where costs are private information, efficient firms will consider themselves able to compete effectively in prices. Restricting price advertising limits their ability to compete using price. In itself this may reduce the profits of efficient firms since customers will tend to be lost to less efficient competitors. In addition, these firms may respond aggressively by cutting price to regain the customers they have lost. If this is true, the equilibrium prices of low-cost firms may actually fall after imposition of the advertising restrictions.

To formalize these arguments, an industry where all firms have limited capacity and private information about their production costs is examined. The Bayesian Nash equilibrium that will prevail in this industry with and without advertising restrictions is computed. In this framework, the question I wish to examine is whether advertising restrictions might provide an indirect means of coordinating the pricing decisions of firms in the industry. Though direct agreements to raise price might be difficult to monitor, and therefore difficult to enforce, violations of a no-advertising agreement can be easily observed. This is the reason that indirect restrictions on advertising might be preferred to direct restrictions on prices.

In Section II, I compute Bayesian Nash equilibria using a technique suggested by Riley and Samuelson (1981) in their study
design. I show that equilibrium pricing strategies ha
that very efficient producers will be hurt by advertisin
the sense that their expected payoffs might fall. I
shown that equilibrium prices of efficient producers
with advertising restrictions than they are without

In Section III, I study the properties of a limiting
the number of buyers and sellers is very large. In this
uncertainty disappears. The equilibrium with advertising collapses to a single price charged by all sellers who stay in the market. This price can be computed by finding the intersection of the demand and supply curve, where the elasticity of the supply curve results from the fact that less efficient sellers are attracted to the market as prices rise. Without advertising, there is a unique two-price equilibrium. This equilibrium involves some firms charging the monopoly price and catering to high-search-cost buyers. Other firms in the economy sell only to low-search-cost buyers and charge a price that depends on the number of buyers who purchase at that price. It is proved that this

price may be below the price they would have charged if advertising were permitted. In addition, it is simple to construct cases where average industry price falls after advertising restrictions have been imposed.

The results I derive in the limiting case are similar to the results of Salop and Stiglitz (1977) on price dispersion. The primary difference between their model and the model presented here is that explicit capacity constraints are imposed here and firm costs are allowed to differ. Cost differences are necessary for it to be possible for the equilibrium price of any firm with advertising to be higher than the corresponding price without advertising. In addition, the fact that costs may differ and may be unobservable makes it possible to compute equilibrium in the case where the number of firms is finite. This case is of clear importance for problems involving cooperative agreements to restrict competition.

II. The Model

A simple industry is considered where J potential sellers can produce a single unit of a homogeneous nonstorable commodity at some constant marginal cost x. Production costs are not publicly known, perhaps because costs of different sellers vary over time. I assume that these costs vary between zero and one. Further, it is assumed that everyone in the economy believes each seller's costs to be identically and independently distributed according to some continuously differentiable distribution function $G(x)$. Sellers are all risk-neutral expected profit maximizers who compete with prices.

There is a finite number I of potential buyers of the industry's output. It is assumed that $I < J$. To be simple, it will be assumed that each buyer's demand curve is horizontal at a price equal to one, up to one unit of output, at which point it becomes vertical. In other words, each buyer is willing to purchase one and only one unit of output at any price less than or equal to one. Resale is prohibited, and indirect utility is assumed given by $1 - p$, where p is the price the buyer pays. If many buyers visit the same seller, it is assumed that the seller selects one of them at random and provides this single buyer with all his output.

The economy is supposed to work in the following way: At the beginning of the trading day, sellers (costlessly) simultaneously advertise their locations, their willingness to trade, and possibly their price. If a price is announced, sellers are bound to sell at that price and cannot make subsequent changes. Buyers can costlessly arrange transactions with sellers who advertise prices, perhaps by using the telephone. Sellers who do not advertise prices must be visited. Buyers may be either high-search-cost buyers, in which case they can visit one and only one seller (when prices are not advertised), or low-search-cost buyers, in which case they search out the best deal available.

Two cases are considered. In the first, all sellers are supposed to advertise prices, while in the second a producers' association is able to prevent all price advertising. Firms still advertise their locations and willingness to trade. In this case a firm price offer is made only after the buyer turns up at the seller's store. It is easy to verify that equilibria where some sellers advertise and some do not can be ignored when the number of buyers and sellers is finite. To see this, consider the lowest price at which any firm would choose not to advertise. The firm charging this price can never be hurt by choosing to advertise, since by doing so it would attract any high-search-cost buyers who would otherwise have gone to other nonadvertising firms. There would be no gain to advertising only in the case where all nonadvertising firms charge the same price. But then there is no loss in generality associated with assuming that firms advertise, since buyers are assumed to know firms' pricing strategies.

To prove the results, I compute Bayesian equilibria using the techniques of Riley and Samuelson (1981). Each seller chooses a strategy rule $\sigma(x)$ that determines the price he will charge when his costs are x. The rule chosen must maximize expected profits for each value x given the strategy rules employed by the other sellers. When sellers advertise, or compete for low-search-cost buyers, only the lowest quoted prices end up attracting buyers. Hence this problem is formally similar to an auction (Ortega-Reichert 1968, Myerson 1981, Milgrom and Weber 1982) where the Bayesian equilibrium concept has been widely employed.

The case where sellers advertise prices is considere [illegible]
pute the equilibrium, suppose that each seller empl [illegible]
egy rule $\sigma(x)$ and that this rule is a continuously [illegible]
strictly monotonically increasing function of the sel [illegible]
later verify that the equilibrium pricing strategies [illegible]
Suppose there are B buyers in all. A seller sells [illegible]
only if he charges one of the B lowest prices. Since [illegible]
pricing rules are monotonic, by assumption, this means that [illegible]
sells if and only if he has one of the B lowest production costs. If the seller has costs x, there is a probability $G(x)$ that any other seller has lower costs. The probability that $B - 1$ or fewer of the $J - 1$ other sellers have lower costs is then given by the cumulative binomial distribution. This gives

$$Q_B(x) \equiv \sum_{n=0}^{B-1} \frac{(J-1)!}{n!(J-1-n)!} G(x)^n [1 - G(x)]^{J-1-n}. \qquad (1)$$

where $Q_B(x)$ is, then, the probability in equilibrium that a seller with cost x will make a sale if all sellers are using the pricing rule $\sigma(x)$. This is a differentiable function provided $G(x)$ is. Note especially that this probability can be easily recomputed for any value of B. This property will be used later in computing strategies when there is no advertising.

Now let $P(x)$ be the expected price received by a seller of type x when he uses his optimal strategy in the Bayesian equilibrium. From the argument above, it must be that no seller can benefit by acting as if his cost is different from what it actually is. Since $P(x) = \sigma(x)Q_I(x)$ and $\sigma(x)$ is assumed differentiable, it follows from (2) that $P'(x) - xQ_I'(x) = 0$ for each x in $[0, 1]$. Since the function $Q_I(x)$ is known, this gives a simple differential equation to be solved for the function $P(x)$. The required boundary condition is given by noting that $\lim_{x \to 1} P(x) = \lim_{x \to 1} \sigma(x)Q(x) = 0$ since it never pays to report a price above one. Integrating and using the boundary condition gives

$$P(x) = xQ_I(x) + \int_x^1 Q_I(t)dt$$

so that

$$\sigma(x) = x + \int_x^1 \frac{Q_I(t)}{Q_I(x)}\, dt. \qquad (2)$$

The expected payoff to the seller is then just $\int_x^1 Q_I(t)dt$.

This will give the solution to the problem provided the equilibrium strategies $\sigma(x)$ are monotonic and differentiable. They are clearly differentiable. To see that they are strictly monotonic, note that $Q_I(x)$ gives the probability that the reservation price x is among the I lowest reservation prices. Then $1 - Q_I(x)$ is just the probability that the Ith-order statistic among the $J - 1$ reservation prices of the other sellers is less than or equal to x. The density for the Ith-order statistic is

$$\frac{(J - 1)!}{(I - 1)!(J - 1 - I)!} G^{I-1}(x)[1 - G(x)]^{J-I-1} \cdot G'(x),$$

which is strictly positive since $G' > 0$. Then $d(1 - Q_I)/dx = -Q_I'(x) > 0$ or $Q_I'(x) < 0$. Straightforward differentiation of (2) then gives the result that $\sigma(x)$ is monotone increasing. Furthermore, since $P(x) - xQ_I(x) \equiv \int_x^1 Q_I(t)dt$, we have $P''(x) - xQ_I''(x) = Q_I'(x) < 0$ so that the appropriate second-order conditions hold.

Now we consider the outcome when sellers force buyers to search out price quotations. The procedure here is identical to that of the last section. We derive the transaction probability for a seller of type x in equilibrium under the assumption that the pricing strategies of all sellers are monotonic. We then compare the equilibria.

When sellers refuse to convey price information, the model unfolds as follows: sellers post information about their locations at the beginning of the trading day and tell buyers that they will *offer* a single "take it or leave it" price to buyers once they visit. The zero-search-cost buyers are then assumed to arrange transactions with all the lowest-priced sellers. The high-search-cost buyers pick among the set of sellers at random and buy if the seller they choose has not made a transaction. The seller's transaction probability then depends on the number of zero-search-cost buyers on the market. If there are B buyers with zero search costs the seller sells if he is among the B sellers with the lowest reservation price or if either of the other $I - B$ buyers chooses him first. Taking expectations over the number of buyers with zero search costs gives the transactions probability as

$$S(x) = \sum_{t=0}^{I} \left\{Q_t(x) + [1 - Q_t(x)]\left(\frac{I-t}{J}\right)\right\} \cdot q_t, \tag{3}$$

where q_t is the probability that there are t low-search-cost buyers, and Q_t, defined by (1), is the probability that the seller is among the t lowest-priced sellers. Set $Q_0 \equiv 0$.

The argument in the last section applies directly to this case. A seller who has reservation value $= 1$ will never accept a price below one. He never sets a price above one; hence the only strategy available to him is his reservation price. His expected price is then $S(1) = \Sigma_{t=0}^{I} q_t[(I - t)/J]$. Then $P_S(1) = 1 \cdot S(1) - \int_0^1 S(t)dt + \text{constant} = S(1)$ so that the appropriate constant is again $\int_0^1 S(t)dt$. The solution for the equilibrium pricing strategy is then

$$\sigma_S(x) = x + \int_x^1 \frac{S(t)dt}{S(x)}.$$

Once again $\sigma_S(x)$ is differentiable. To see that it is that $S(x)$ can be written as

$$S(x) = \sum_{t=0}^{I} q_t\left[\left(\frac{I-t}{J}\right) + \left(1 - \frac{I-t}{J}\right)Q_t(x)\right],$$

which is monotonic because $Q_t(x)$ is.

Unfortunately, $S(x)$ and $Q(x)$ do not yield simple comparisons. However, we do have the following proposition.

Proposition 1: There exist some sellers with very high costs who are always better off in the equilibrium without advertising.

Proof: Since the value of the game with search is $\int_x^1 S(t)dt$, while without search it is $\int_x^1 Q_I(t)dt$, the payoffs can be computed by comparing S and Q_I. If x goes to one, $Q_I(x)$ goes to zero while $S(x)$ goes to $\Sigma_{t=0}^{I} q^t[(I - t)/J] > 0$. By continuity, $S(x) > Q_I(x)$ in some neighborhood of one. Hence for large enough x, $\int_x^1 S(t)dt > \int_x^1 Q_I(t)dt$. Q.E.D.

Unfortunately, this is a rather weak proposition because it says little about the preferences of more efficient sellers. In the Appendix, a number of properties of the functions $S(x)$ and $Q_l(x)$ are examined. These properties allow us to establish the following.

Proposition 2: The functions $\int_x^1 S(t)dt$ and $\int_x^1 Q_l(x)dt$ are equal at, at most, one point x^*. Further,

$$\int_x^1 S(t)dt \geq \int_x^1 Q_l(t)dt, \quad x \geq x^*,$$

while

$$\int_x^1 S(t)dt \leq \int_x^1 Q_l(t)dt, \quad x \leq x^*.$$

This proposition justifies the intuitive categorization of sellers into high- and low-cost firms. The proposition states that firms with costs above x^* are made better off by advertising restrictions while those with costs below x^* are made worse off.

Unfortunately, there is no way to tell whether $x^* > 0$. It may be equal to zero. However, we do have:

Proposition 3: If $\int_x^1 S(t)dt < \int_x^1 Q_l(t)dt$ for some x, then there is an open set of very efficient sellers whose prices fall when advertising is restricted.

Proof: The seller with the lowest cost ($x = 0$) sets a price $\sigma_S(0) = \int_0^1 S(t)dt$ and $\sigma(0) = \int_0^1 Q_l(t)dt$. The result then follows by the continuity of the equilibrium pricing strategies. Q.E.D.

It is easy at this point to construct an example to show the desired result.

For this example, let there be two buyers and three sellers and suppose that $G(x) = x$. From (1), $Q_2(x)$ is given by $(1 - x)^2 + 2x(1 - x)$. Then $\int_x^1 (1 - t)^2 + 2t(1 - t)dt = [(1 - x)^3/3] + (1 - x^2) - \frac{2}{3}(1 - x^3)$, which varies from $\frac{2}{3}$ when $x = 0$ to zero when $x = 1$. From (2), $S(x) = q^0\frac{2}{3} + q^1[\frac{1}{3} + \frac{2}{3}(1 - x)^2] + q^2[(1 - x)^2 + 2x(1 - x)]$, so that $\int_x^1 S(x) = q^0\frac{2}{3}(1 - x) + q^1\{\frac{1}{3}(1 - x) + \frac{2}{3}[(1 - x)^3/3]\} + q^2\{[(1 - x)^3/3] + (1 - x^2) - \frac{2}{3}(1 - x^3)\}$. When x is zero this is $\frac{2}{3}(q^0 + q^2) + q^1(\frac{1}{3} + \frac{2}{9})$. This is always less than $\frac{2}{3}$ provided q^1 is nonzero. It then follows that sellers with low reservation prices are always worse off without advertising. It is straightforward to show that if there is one buyer and two sellers, all sellers are better off without advertising, so that examples can be constructed in either direction.

What about prices? If we let the probability that there are no high-search-cost buyers be zero, then $S(0)$, the probability of selling when costs are zero, goes to one. We then have $\sigma_S(0) \cdot S(0) = S(0) \cdot 0 + \int_0^1 S(t)dt = \sigma_S(0) = q'(\frac{2}{3}) + q'(\frac{1}{3} + \frac{2}{9}) < \frac{2}{3}$ for, say, $q = \frac{1}{2}$. When there is advertising, a seller with zero costs always sets price equal to $\frac{2}{3}$.

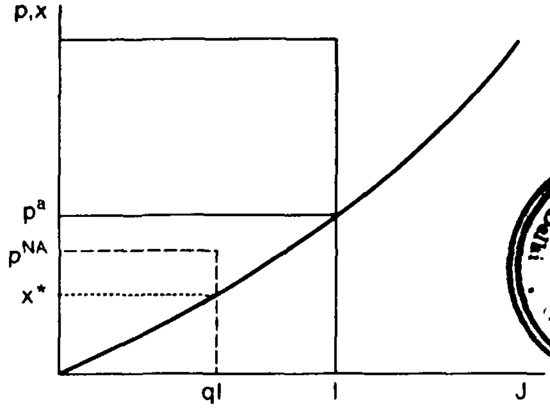


FIG. 1

III. Many Buyers and Sellers

When the number of buyers and sellers is large, the analysis of this model is greatly simplified, though the relevance of cooperative restrictions on advertising can be questioned. In the limit, the proportion of sellers having costs less than x is $G(x)$ with certainty. Further, q becomes the proportion of buyers having zero search costs.

It will now be shown that the equilibrium pricing strategies tend in the limit toward piecewise constant functions. The reason for this is basically that when prices are advertised and J is large, a seller is able to tell with certainty whether or not he will make a sale. If his
x, then he knows that the proportion $G(x)$ of all se
costs. A sale will be made if and only if $G(x) \cdot J \leq I$. I
that $G(x) \cdot J \leq I$, no increase in selling probability c
cutting price because of the capacity constraint fac
all sellers with costs such that $G(x) \cdot J \leq I$ will set the
price, p^a, will be equal to the cost of the seller for w
This price can be thought of as a competitive price
the supply curve, while the demand curve is vertical at I

In the case where prices are unadvertised, sellers can still compete for the qI buyers who have zero search costs. In this case no single-price equilibrium can exist since shading price guarantees a sale to any seller who otherwise would have sold with probability I/J (actually all sales are certain, at the single price each seller sells I/J units). The two-price equilibrium that applies in this case can be constructed in the following way. Choose the qI lowest-priced sellers. Let the cost of the highest-cost seller in this group be x^*. These sellers will all set a single price $p^{NA} > x^*$. The other $(1 - q)I$ high-search-cost buyers are

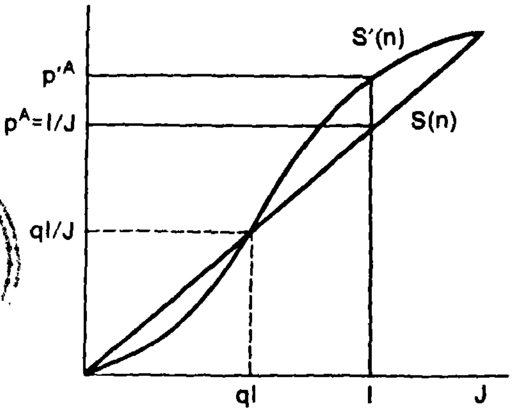


FIG. 2

distributed evenly among all of the existing sellers. Some of these buyers will visit low-cost firms and make no transaction. The equilibrium price p^{NA} is chosen such that the seller with cost x^* is indifferent between selling 1 unit at price p^{NA} and selling $(1 - q)I/J$ units at price 1; that is, p^{NA} is chosen to satisfy

$$(p^{NA} - x^*) = (1 - x^*)\frac{(1 - q)I}{J}. \tag{4}$$

It is easy to check that firms with costs below x^* prefer the price p^{NA}, while firms with costs above x^* prefer price 1. It is also straightforward to verify that this equilibrium is unique.

As in the finite case, we must check whether the lower price in the two-price equilibrium without advertising might actually be below the single competitive price that is charged with advertising. This depends on the distribution function $G(x)$. To see this, suppose that the distribution function $G(x)$ is uniform, so that the "supply" curve is a straight line with slope I/J. The price with advertising is then I/J, while the cost of the marginal seller in the two-price equilibrium is qI/J (see fig. 2).

The price p^{NA} must be chosen so that the rectangle with height $p^{NA} - (qI/J)$ and length J has area equal to the rectangle with height $1 - (qI/J)$ and length $(1 - q)I$.

It is easy to check that the lower price without advertising will be below the single advertised price in this case. Solving (4) when $G(x)$ is uniform yields $p^{NA} = (I/J) - [(1 - q)qI/J^2] < I/J$.

Now, suppose we change the supply function in a way that makes any new supply curve pass through the point $(qI/J, qI)$ on the linear supply. This could be done, for example, by taking mean-preserving spreads of the distribution $G(x)$. Such a mean-preserving spread would shift the supply curve to a new curve like $S'(n)$ in figure 2. This spread would not change the equilibrium price p^{NA}, since the equilibrium condition would continue to hold. It would, however, affect the equilibrium price that would prevail with advertising. Mean-preserving spreads of $G(x)$, like the one demonstrated in figure 2, would drive up the equilibrium price p^A to p'^A.

This procedure makes it clear that a somewhat stronger result is possible. Extreme spreads of the distribution of $G(x)$ can be taken to make the equilibrium price with advertising as close as desired to one. The limiting supply curve would be flat at zero up to ql, at which point it would jump to one, making the equilibrium price with advertising equal to one. The average price charged by firms without advertising is $(ql/J) \cdot p^{NA} + [(J - ql)/J] \cdot 1$, which is strictly less than one. This shows that it is possible that advertising restrictions might cause *average* prices in an industry to fall.

To get some insight into when this might occur note the results that might occur at the other extreme. Reducing the spread in firms' costs leads in the limit to a situation where all costs are the same. Suppose all firms have costs ql/J, so that the supply curve is flat at that level. The price with advertising is ql/J and all rents go away. In this case firms in the industry much prefer advertising restrictions since sales are always made at higher prices. From these arguments it appears to be true that dispersion of costs in an industry is likely [illegible] tising restrictions to bring down average industry [illegible]

IV. Conclusion

This paper analyzes the effects that prohibition o[illegible] has on equilibrium in an industry producing an [illegible] product. It is shown that average prices charged by [illegible] response to this restriction. This conclusion is derived by [illegible] the cost side of the usual argument. The results illustrate a much broader point, that is, that restrictions on competition will tend to hurt efficient firms and benefit inefficient ones. Explicit examination of cost differences makes it possible to demonstrate this proposition here.

The paper suggests a couple of issues that should be pursued further. First, I have severely limited the advertising strategies available to firms in the industry. This serves to make the no-advertising equilibrium quite unstable in the sense that high-priced firms have an incentive to advertise a slightly lower price than their competitors, thereby increasing their sales. The usual recourse to punishment strategies of members of the industry applies. However, no-advertising agreements can be made self-policing in other ways. For example, firms might advertise a promise (legally binding) to match any advertised price offer. Accepting the obvious feasibility problems, this would seem to make advertising strategies by high-priced firms less attractive. An analysis of price advertising with broader strategy sets of this sort would be in order.

Second, I have assumed that the commodity traded is homogeneous. One of the arguments for prohibiting price advertising might be that the resulting price competition would reduce firms' expenditures on quality below a socially optimal level. An additional difficulty arises, however: consumers would be forced to make inferences about quality from prices or from the decision not to advertise. The range of possible outcomes seems large.

Advertisements in this model convey more than information about price. They convey the firms' willingness to utilize an efficient informational technology to conduct transactions. Buyers actually search less efficiently among nonadvertising firms. It would be useful to extend the analysis to the case where messages conveyed by firms were pure information transmitted at some cost.

Finally, the possibility of entry into this industry has been ignored. The reason for this, again, is that in an industry where cooperative agreements to restrict advertising are feasible, one should expect the industry to have some control over entry. The issue of entry is important, however, since advertising restrictions are often imposed by lawmakers on essentially competitive industries.

Appendix

Lemma A1: For every $B = 1, \dots, J$,

$$-Q_B^1(x) = \frac{(J-1)!}{(B-1)!(J-B)!} G(x)^{B-1}[1 - G(x)]^{J-1-B} \cdot \frac{dG(x)}{dx},$$

where $Q_B^1(x)$ is the first derivative of $Q_B(x)$.

Proof: $Q_B(x)$ is the probability that $t - 1$ or fewer of the remaining $J - 1$ sellers will have reservation prices below x:

$$Q_B(x) = \sum_{t=1}^{B} \frac{(J-1)!}{(t-1)!(J-t)!} G(x)^{t-1}[1 - G(x)]^{J-t}.$$

This is equal to the probability that a seller of type x makes a sale when there are B buyers on the market. This is equal to the probability that the Bth-order statistic for reservation prices of the $J - 1$ other sellers is not less than x, which is one minus the probability that the Bth-order statistic is less than or equal to x. The density for the Bth-order statistic is

$$\delta_B(x) = \frac{(J-1)!}{(B-1)!(J-1-B)!} G(x)^{B-1}[1 - G(x)]^{J-1-B} \cdot \frac{dG(x)}{dx},$$

which gives the result. Q.E.D.

Lemma A2:

$$S^1(x) < 0, \quad Q_J^1(x) < 0.$$

Proof: The results follow from lemma A1 and the assumption that $G(x)$ is monotone increasing and differentiable.

$$S(x) = \sum_{t=0}^{I} q^t \left\{ Q_t(x) + [1 - Q_t(x)] \frac{I - t}{J} \right\},$$

with $Q_0(x) \equiv 0 \equiv Q_0^1(x)$. Whence

$$S^1(x) = \sum_{t=0}^{I} q^t \left[\frac{J - I + t}{J} Q_t^1(x) \right],$$

so that $S^1(x) < 0$. Q.E.D.

Lemma A3: The functions $S^1(x)$ and $Q_J^1(x)$ have the following properties: (i) $|S^1(x)| > |Q_J^1(x)|$ for all x in an open neighborhood of zero (ii) $|S^1(x)| < |Q_J^1(x)|$ for all x in an open neighborhood of one (iii) There exists a unique x^0 such that $|S^1(x^0)| = |Q_J^1(x^0)|$.

Proof: We compute

$$|S^1(x)| - |Q_J^1(x)| = \sum_{t=0}^{I} q^t \left[\frac{J - I + t}{J} |Q_t^1(x)| - |Q_J^1(x)| \right]$$

$$= \sum_{t=0}^{I} q^t \left(\frac{J - I + t}{J} \left\{ \frac{(J-1)!}{(t-1)!(J-1-t)!} G(x)^{t-1}[1 - G(x)] \right\}^{J-1-t} \frac{dG(x)}{dx} \right.$$

$$\left. - \frac{(J-1)!}{(I-1)!(J-1-I)} G(x)^{I-1}[1 - G(x)]^{J-1-I} \frac{dG(x)}{dx} \right)$$

$$= \sum_{t=0}^{I} q^t \frac{(J-1)!}{J} \frac{dG(x)}{dx} G(x)^{I-1}[1 - G(x)]^{J-I-1}$$

$$\cdot \left\{ \frac{(J - I + t)G(x)^{t-I}[1 - G(x)]^{I-t}}{(t-1)!(J-1-t)!} - \frac{J}{(I-1)!(J} \right.$$

If x is very close to zero, every term in this expansion is p
$- |Q_B^1(x)| > 0$, whereas if x is close to one, every term in this
negative, making $|S^1(x)| - |Q_B^1(x)| < 0$. The sign of this expression is determined by the sign of the expression

$$\sum_{t=0}^{I} q^t \left[\frac{H(x)^{I-t}(J - I + t)}{(t-1)!(J-1-t)!} - \frac{J}{(I-1)!(J-1-I)!} \right]$$

$$= \sum_{t=0}^{I} q^t \frac{H(x)^{I-t}(J - I + t)}{(t-1)!(J-1-t)!} - \frac{J}{(I-1)!(J-1-I)!},$$

where $H(x) = [1 - G(x)]/G(x)$ is a monotonically decreasing function. Hence

$$\sum_{t=0}^{I} q^t \frac{H(x)^{I-t}(J - I - t)}{(t-1)!(J-1-t)!}$$

is a monotonically decreasing function that intersects the constant function $J/[(I-1)!(J-1-I)!]$ only once. Hence the expression changes sign exactly once.

Lemma A4: The functions $S(x)$ and $Q_I(x)$ have the properties (i) $S(x) < Q_I(x)$ for all x near zero. (ii) $S(x) > Q_I(x)$ for all x near one. (iii) There is a unique x^0 such that $S(x^0) = Q_I(x^0)$.

Proof: $\lim_{x\to 0} Q_I(x) = 1$, since a seller with the lowest possible reservation price is almost sure to win.

$$\lim_{x\to 0} S(x) = q^0 \frac{I}{J} + (1-q^0)1 < 1.$$

This is because there may be some possibility that all sellers have high search costs. In this event, the seller with the lowest reservation price has the same transaction probability as everyone else, that is, I/J, which is less than one by assumption. By continuity, condition i holds. Now $\lim_{x\to 1} Q_I(x) = 0$ since $I < J$ by assumption. However, $\lim_{x\to 1} S(x) = \Sigma_{i=0}^{B} q_i[(I-i)/J] > 0$. By continuity, condition ii holds.

Continuity implies that $S(x)$ must cross $Q_I(x)$ at least once. At the largest crossing point, $S(x)$ must be flatter than $Q_I(x)$, since $S(x)$ must eventually be larger than $Q_I(x)$. If it crosses more than once, it must cross an odd number of times. Let $\{x_1, x_2, \ldots, x_I\}$ be the sequence of crossing points. Then $S^1(x_1) > Q_I^1(x_1)$, $S^1(x_2) < Q_I^1(x_2)$ . . . $S^1(x_I) < Q_I^1(x_I)$. But this contradicts the single crossing property of $S^1(x)$ and $Q_I^1(x)$ given in lemma A3. Q.E.D.

Lemma A5: $\int_x^1 S(t)dt \geq \int_x^1 Q_I(t)dt$ for all x or there exists a unique x^0 such that

$$\int_x^1 S(t)dt \leq \int_x^1 Q_I(t)dt \quad \forall\, x \leq x^0,$$

$$\int_x^1 S(t)dt \geq \int_x^1 Q_I(t)dt \quad \forall\, x \geq x^0.$$

Proof. Define the function $\hat{S}(x) = \int_x^1 S(t)dt$ and $\hat{Q}(x) = \int_x^1 Q_I(x)dt$. These are positive functions with derivatives given by $-S(x)$ and $-Q_B(x)$, respectively. If the functions cross more than once they must cross an odd number of times, but this contradicts the single crossing property of the derivative functions given in lemma A4. Q.E.D

Lemma A6: Let x be any point at which $\int_x^1 S(t)dt \geq \int_x^1 Q_I(t)dt$. Then $\int_x^1 S(t)dt/S(x) > \int_x^1 Q_I(t)dt/Q_I(x)$.

Proof: If $S(x) < Q_I(x)$ the proof is immediate, so suppose that $S(x) > Q_I(x)$. By lemmas A3 and A4 $S(t)$ is flatter than $Q_I(t)$ for $t \geq x$. Now construct the function $\bar{S}(t) = S(t) - [S(x) - Q_I(x)]$. $\bar{S}(t)$ is the function that is shifted downward vertically until $\bar{S}(x) = Q_I(x)$. Since $\bar{S}^1(t) = S^1(t)$, $\bar{S}(t)$ is flatter than $Q_I(t)$ for all $t \geq x$. Hence $\int_x^1 \bar{S}(t)dt > \int_x^1 Q_I(t)dt$. Let λ be the positive real number having the property that $\lambda\bar{S}(x) = S(x)$. Note that because S is a decreasing function, $\lambda\bar{S}(x) \leq S(x)$ (see fig. 1). Hence

$$\frac{\int_x^1 S(t)dt}{S(x)} > \frac{\int_x^1 \lambda\bar{S}(t)dt}{\lambda\bar{S}(x)} > \frac{\int_x^1 Q_I(t)dt}{Q_I(x)}. \text{ Q.E.D.}$$

Proposition 3 follows from lemma A5, while proposition 2 follows from lemma A6.

References

Benham, Lee. "The Effect of Advertising on the Price of Eyeglasses." *J. Law and Econ.* 15 (October 1972): 337–52.

Boyer, Kenneth D. "Informative and Goodwill Advertising" *Rev. Econ. and Statis.* 56 (November 1974): 541–48.

Milgrom, Paul R., and Weber, Robert J. "A Theory of Auctions and Competitive Bidding." *Econometrica* 50 (September 1982): 1089–1122.

Myerson, Roger. "Optimal Auction Design." *Mathematical Operations Res.* 6 (February 1981): 58–73.

Ortega-Reichert, A. "Models for Competitive Bidding under Uncertainty." Ph.D. dissertation, Stanford Univ., 1968.

Riley, John G., and Samuelson, William F. "Optimal Auctions." *A.E.R.* 71 (June 1981): 381–92.

Salop, Steven, and Stiglitz, Joseph E. "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion." *Rev. Econ Studies* 44 (October 1977): 493–510.

Work Incentives, Hierarchy, and Internal Labor Markets

James M. Malcomson

University of York

This paper argues that contracts with payment based on a ranking of employee performance can provide performance incentives even under asymmetric information that prevents payment based on individual performance only being enforceable. Such contracts also fit with five features of labor markets that have aroused considerable interest: (1) hierarchical wage structures; (2) internal promotion; (3) wage rates that rise with seniority and experience more than productivity; (4) the variance of earnings increasing with experience; and (5) wage rates attached to jobs rather than individuals with differentials set by administrative procedures rather than by reference to external market wages.

Introduction

The literature on the principal-agent problem has been concerned with the nature of optimal contracts when there is difficulty in monitoring all aspects of an agent's performance. For the most part, this literature has considered models in which payment based directly on a measure of output provides an enforceable contract. For a recent example, see Grossman and Hart (1983). This requires that information about an agent's output be symmetric so that both parties know what payment is due under the contract and, at least in principle,

---

I wish to thank Arnold Arthurs, Lorne Carmichael, Oliver Hart, Keith Hartley, Ravi Kanbur, Peter Lambert, Andrew Oswald, Robert Solow, an anonymous referee, and participants in numerous seminars for valuable comments on earlier versions of this paper. They are, of course, none of them responsible for any remaining errors.

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0003$01.50

could get that payment enforced legally if necessary. The output measure does not have to be perfect as long as neither party is in a position to impart a systematic bias to it. However, in many employment situations, information about an agent's output is asymmetric in the sense that the agent cannot verify the principal's observation of it. As Greenwald (1979, p. 8) and Williamson, Wachter, and Harris (1975) have emphasized, performance is frequently something that can be assessed only subjectively by, for example, the judgment of a supervisor. This is particularly the case where there is a genuine team element in production of the type that Alchian and Demsetz (1972) argue is a fundamental reason for the existence of firm, as opposed to purely market, organization of production. Under such circumstances a payment scheme based directly on a measure of individual performance ("payment by standard," as it has become known in the literature) is unenforceable, and, unless some other kind of incentive scheme is available, the principal will be unable to enforce more than minimal compliance by the agent in the performance of his task.[1]

It is the contention of the present paper that in the context of a principal employing many agents, which characterizes much employment by firms, contracts with payment based on an ordering of employee performance, and hence in the spirit of the rank-order tournaments of Lazear and Rosen (1981), can still be made enforceable as a means of providing performance incentives even under asymmetric information. This provides a wider set of circumstances than those analyzed by Green and Stokey (1983), that is, situations in which stochastic elements unobservable by the principal are
across agents, under which such tournaments are
ment schemes based on individual performance o

The essence of the argument is that a firm ca
specifying that a certain proportion of its work fo
higher wage than the rest. Whether or not it abid
thing that can be verified by employees. Since the
mitted to paying a higher wage to a certain proportion
it has every incentive to pay that higher wage to those emplo

---

[1] Piece-rate payment is one such form of incentive scheme. If the basis for piece-rate payment can be the true marginal product of an employee, then the situation essentially reverts to a principal-agent problem in which the agent's output is verifiable by both parties. The problem with piece rates is that the unit of output used as the basis of payment must be verifiable by both parties at reasonably low cost and hence tends to be a quantitative measure that neglects all but the crudest quality dimensions. As Stiglitz (1975) has observed, "because piece rates reward speed, they tend not to provide the correct incentives for quality or proper care of equipment." Under such circumstances, even with piece-rate payment, it may be possible to improve efficiency with the use of schemes of the type that form the focus of this paper. Note also that profit-sharing bonuses are not an adequate substitute in a large enterprise because they do not directly reward individual effort.

perform best because this provides the greatest incentive for employees to perform well. But this type of contract amounts to a contract with payment based on an ordering of employees' performance, since it rewards the best no matter what the overall level is. Moreover, because a promotion scheme that ensures that a given proportion of employees is promoted exactly fits the requirements of such a contract, reward schemes of this type would seem to be important in practice.

In a repeated game of the principal-agent type, it may not be necessary to have full legal enforceability of a contract to achieve de facto enforceability, as has been shown by Radner (1981). If the principal and agent have a very long relationship with each other, then an equilibrium in which neither defaults may exist even when it pays to default in the short term. The essential reason is that a no-default equilibrium can benefit both parties in the long term, so that, with a long enough time horizon and a low enough discount rate, the cost to the principal of losing a reputation for honesty, and hence not being trusted not to default in the future, may exceed the short-term gain from defaulting. Such an equilibrium may not, however, exist and is never the only equilibrium since, when a no-default equilibrium exists, there always exists a Nash equilibrium based on the assumption that the principal will default. Hence such an equilibrium, even when it exists, may well not prove robust. The use of a contract basing reward on an ordering of employee performance is, however, always available when there are many agents and can produce efficiency gains without relying on anything more than Nash strategies.

The use of such contracts also fits with certain features of employment arrangements, particularly in large organizations, which have aroused considerable interest. These are:

1. The wage structure within an organization is hierarchical in the sense that there is a tendency for the number of employees being paid higher wages to be less than the number being paid lower wages. This is what Stiglitz (1975) terms "horizontal hierarchy."

2. A high proportion of those in higher-paid jobs have been promoted from lower-paid jobs within the same organization, and new entrants are for the most part appointed only at specific points in the hierarchy. These are characteristics of the internal labor market structure documented by Doeringer and Piore (1971).

3. Wage rates rise more with seniority and experience than productivity does. This has been documented by Medoff and Abraham (1980). For many employees it is associated with a mandatory retirement age.

4. The variance of earnings increases with experience (Mincer 1974).

5. "[W]age rates attach mainly to jobs rather than to workers" (Williamson et al. 1975). While not entirely unambiguous, the fundamental idea here seems to be that, once assigned to a particular job, there is little scope for an individual to bargain over the rate for that job. Moreover, the wage differentials between jobs are often set more by administrative procedures than by reference to external market wages (see Doeringer and Piore 1971).

One reason for the interest in these features is that they are, for the most part, inconsistent with the simplest notion of employees being paid a wage in every period equal to their marginal product in that period. Moreover, they cannot just be ascribed to collusive action on the part of trade unions since they are not present only in highly unionized markets.

The existing literature is not short of explanations for these features, but many of those explanations are directed to just one and do not fit easily with all the others. It is not the purpose of the present paper to deny the validity or importance of these earlier explanations. Indeed, many of their insights are incorporated here. As will be shown, however, the different features fit together in a natural way when contracts base rewards on an ordering of employee performance.

The literature on feature 1 stems from the work of Simon (1957) and Lydall (1968). Their explanation of hierarchical structure is somewhat mechanical since it assumes that a supervisor can supervise several employees and is always paid a wage higher th
supervises. Calvo and Wellisz (1979) provide an ex
of the optimal monitoring structure when increas
proves work performance by increasing the prob
shirking. The higher wage for supervisors in the
follows from the somewhat special assumption th
individuals cannot be distinguished by the firm
before they are hired, the firm knows that they have a m
tion wage.

Feature 2 arises naturally in any model that involves firms' acquiring better information about an employee's ability on the job than can be acquired from characteristics apparent at the time of hiring and information available from previous employers. Then early periods of employment act as a screening process. The wage offered to new entrants to the firm therefore will reflect the mixed ability of newly hired employees, and the more highly paid will be those who have worked for the firm previously and have been revealed to have greater ability. This would also generate feature 4. Explanations of this type have been offered by Williamson et al. (1975), Greenwald (1979), and Harris and Holmstrom (1982). In Greenwald's model of adverse selection, however, wages fall relative to the expected productivity of an individual, which does not fit with feature 3, and Harris and Holmstrom's result on this fits with feature 3 only in a rather special sense. Moreover, to the extent that feature 2 involves promotion to a different job or grade, it is only necessary as a means of rewarding higher productivity if feature 5 holds. This feature is ascribed by Williamson et al. (1975) to an attempt to reduce bargaining costs in the face of what they call idiosyncratic exchange—that is, when job-specific skills give individual employees a measure of monopoly power in wage bargaining—but no formal model of this has been offered.

The existence of job-specific skills acquired by experience forms the basis of Carmichael's (1983) explanation of features 2 and 3. In his model, a fixed number of employees who have acquired the skills are offered a higher wage than their equally skilled colleagues on the basis of seniority in order to reduce quitting by the skilled. This has better efficiency properties than a separation penalty paid by the party initiating the separation to the other party. The reason is that it provides no incentive for the party wishing to separate to act in such a way as to avoid incurring the penalty by inducing the *other* party to initiate the separation. The results of the present paper, however, make it clear that the existence of job-specific skills is not a necessary condition for the type of employment arrangement characterized by features 2 and 3 to emerge.

Holmstrom (1981) generates feature 3 as a result of the use of state-contingent implicit contracts that allow risk-neutral firms to bear some of the risk that would otherwise be borne by risk-averse employees. In common with many other implicit contract models, however, this model uses the crucial assumption that firms are not allowed to pay severance payments to employees. The explanation for this feature closest in spirit to the model used here is that of Lazear (1979, 1981), which can be outlined as follows. To encourage employees not to shirk, employers pay less than marginal product in the early years of employment and more than marginal product in later years. This increases the costs to an employee of being dismissed for shirking, as a result of which less shirking occurs and a Pareto improvement is possible. Mandatory retirement occurs at the point at which an individual's reservation wage becomes greater than his marginal product, which is socially efficient. But with the wage above marginal product at this time the individual would not choose to retire so that mandatory retirement needs to be written into the employment contract.

Two features of Lazear's model are not wholly satisfactory. One is that performance is rated either as satisfactory or as unsatisfactory. There are no degrees of satisfactoriness, and this leads to an indeterminacy as to the optimal path as long as the cumulative wage paid by any date is less than the cumulative marginal product. The other is an enforceability problem. In his model a firm can dismiss an employee for shirking even though he did not shirk and so capture the difference between the cumulative wage paid and the cumulative marginal product to the date of dismissal. Then, of course, there can only be a solution with the cumulative wage to any date less than the cumulative marginal product if there is a cost to a firm of defaulting in this way. Otherwise, it would always pay firms to default at the point where the difference between these is at a maximum. Lazear assumes that there is such a cost that arises because a firm that acquires a reputation for defaulting will not be able to use such an incentive arrangement in the future and hence will incur higher labor costs per unit of output. This of course amounts to a repeated game of the type discussed briefly above and for which the existence of a no-default equilibrium is by no means guaranteed. As this paper will show, however, the essential spirit of Lazear's insight can survive the removal of both these special assumptions when contracts basing payment on an ordering of performance of employees are used.

The next section of this paper shows how such contracts can be made enforceable as a means of providing performance incentives even with asymmetric information about employees' performance and also shows that promotion schemes are a way of achieving this. Subsequent sections analyze a formal model of such a scheme, show that it can lead to efficiency improvements over other enforceable types of contract, and demonstrate how it can generate the features of employment arrangements discussed above.

Enforceability of Contracts

To see precisely the implications of asymmetry of in [illegible] employee performance, consider the following simple [illegible] may hire employees of different generations. All individuals [illegible] tical (apart from belonging to different generations) and may work for up to 2 periods (indexed by $j = 0, 1$, respectively, for each generation) so that at any one date individuals of two generations may be employed. Each individual has a utility function $U(e_0, y_0, e_1, y_1)$, where e_j is work effort and y_j is consumption in period j. It is assumed conventionally that $U_1(\cdot), U_3(\cdot), U_{11}(\cdot) < 0$ and $U_2(\cdot), U_4(\cdot) > 0$. Employees maximize expected utility. The firm is perfectly competitive, which, in terms of purchasing labor, means that it takes as given the minimum expected utility level, $\bar{U}$, that its employment contract must offer if it is to hire employees and it believes it can hire as many employees as it wishes at that expected utility level. Also, for expositional simplicity, constant returns to scale are assumed so that output is proportional to effort. This allows the firm to choose the optimal contract for any one generation of employees independently of any other generation. All the results do in fact continue to hold as long as the firm's revenue function is concave, but the proofs become more complicated and no additional insights arise. The firm's revenue from a unit of effort from a period 0 employee will be normalized at one. To allow for experience-related increases in skills, that for a period 1 employee will be denoted s_1. The firm is assumed to be able to borrow and lend on a perfect capital market at an interest rate $(1/r) - 1$.

The firm makes a subjective assessment of each employee's work effort, subject to measurement error. In the spirit of the preceding discussion, it is assumed that this assessment is not verifiable by employees. Let $f(x - e)$ be the probability density function for observed effort x given true effort e. This form makes the implicit assumption that the distribution depends only on the error and not on the level of e, which simplifies the proofs without losing the essential idea. Also, for simplicity, it is assumed that $f(x - e)$ is differentiable and strictly positive for all values of $(x - e)$ strictly between its upper and lower bounds. In what follows, one can either interpret e as the true output of an employee and $(x - e)$ as a measurement error or interpret x as true output related to effort by a random variable with zero mean. In the latter interpretation the expressions given below for the firm's profit are, in fact, just its expected profit. There is one further assumption about work effort. As some level of effort (e.g., whether an employee is actually present on the job) is objectively verifiable at low cost and may yield some output, it will be assumed that the firm can enforce some minimum effort level $\bar{e}$ as a condition of payment.

The variable e can be interpreted in a number of ways. If employees are hired by the day, month, or year, it can represent the hours they actually work. If hired by the hour, it can represent how hard they work during the hour. On an assembly line job, it can represent how well they perform their task, for example, how soon after leaving the factory the component they are responsible for needs attention. Whichever interpretation is appropriate, it is important to note that forms of industrial action such as a "work-to-rule" or a "go-slow" would impose a cost on the firm only if actual performance is in fact above the minimum the firm is able to enforce.

In this context, the firm cannot provide an incentive for work effort greater than $\bar{e}$ by making the wage paid depend on x. Since x cannot be verified by employees, such a payment scheme would be unenforceable. Indeed, the only payment based directly on a standard that is enforceable is a prespecified wage that is paid conditional only on effort of at least $\bar{e}$, but, with $U_1(\cdot) < 0$, this would never induce work effort greater than $\bar{e}$. However, a payment scheme based on ordinal performance can both be made enforceable and provide such an incentive. To see this, note that the firm can offer a contract in which, provided they perform at least at the level $\bar{e}$, employees receive one of two wages, w_1 and w_2, with $w_2 > w_1$, and the higher wage w_2 is paid to a specified proportion P of employees. As long as the proportion P actually paid the higher wage is verifiable by employees, such a contract is enforceable. Moreover, since x and e are positively correlated, the contract provides an incentive for increasing work effort provided the firm makes up the proportion P who are paid the higher wage from those at the top of the distribution of observed effort. The firm, of course, has every incentive to do just that. Since P, w_1, and w_2 are all specified in the contract, it cannot reduce its wage bill by dissembling in any way. Falsifying the ordering of employees' observed effort would merely reduce the incentive to increase effort without reducing its wage bill. Moreover, understating the observed effort of all employees has no effect whatever since P, w_1, and w_2, and hence the firm's costs, are independent of the absolute values of the observed effort levels.[2]

In fact, in the present model specifying P, w_1, and w_2 is precisely equivalent to being able to specify a value of x, say x^*, with all employees having observed effort greater than x^* paid the higher wage and all those having observed effort lower than x^* paid the lower wage, at least as long as the firm is sufficiently large that the sampling variation in the observed values of x is negligible. Then, making up the proportion P from those employees at the top of the observed e [illegible]
tion is equivalent to defining x^* by

$$P = 1 - F(x^* - e),$$

where $F(x - e)$ is the cumulative distribution func [illegible]
effort. Note that, for the contract to provide a gen [illegible]
effort, P must lie in the range $0 < P < 1$ since if [illegible]
everybody is paid the higher wage no incentive effect rem [illegible]
$P = 0$ or $P = 1$, the contract simply reduces to paying a fixed wage conditional on effort at least as great as $\bar{e}$ (i.e., payment by a standard). In fact, of course, there is no need to limit the number of wages on offer to just two. One could have a whole range of possible wage rates with specified proportions of employees being paid each. Since, how-

---

[2] Strictly speaking, if employees had no way of assessing either the true or observed effort levels of other employees, they could not actually tell whether the firm was in fact rewarding effort, so the firm could save itself the trouble of measuring x by determining those to be paid the higher wage at random. Implicitly, then, it is assumed that employees observe enough of their colleagues' behavior eventually to find out if merit goes systematically unrewarded. That does not seem unreasonable

ever, most of the interesting issues arise in the simple case, such complications will be neglected here.

As was suggested above, one particular form of reward that has the characteristics of an ordinal system is a promotion scheme. Instead of offering a 1-period contract with a higher wage going to a specified proportion of employees, the firm could offer a 2-period contract with the same wage w_0 paid to all employees for work in period 0 of their working lives and a specified proportion promoted to a higher-paid grade for period 1 of their working lives. Then P becomes the proportion promoted, w_2 the wage in period 1 for those promoted, and w_1 the wage in period 1 for those not promoted. The incentive characteristics of such a 2-period contract are similar to those for the 1-period contract.

There are, however, good reasons for the use of the 2-period form of contract in practice. The 1-period form would require payment to be made at the end of the period over which work performance is reviewed. Because ranking of employees for payment purposes is not, in general, completely costless to the firm, such reviews will take place only at discrete intervals of time. But employees will want to consume during that period and, since firms typically have cheaper access to capital markets than employees, it will be efficient for firms to finance this consumption by paying in advance of the performance review a wage w_0 that obviously cannot depend on the result of that review. Payment of the performance-related component will then be used to finance consumption in the subsequent period. Since in the present model there is no reason for an employee not to work for the same firm in the second period as in the first (indeed, if firm-specific human capital is at all an important aspect of employment, it would be inefficient for him not to do so), the payment of the performance-related component of the wage can be treated simply as part of the wage for working in the second period. Moreover, since in this model there is no reason for firms to set w_0 higher than the optimal consumption level in the first period, consumption in each period will be the same as income in each period. This is what is assumed in what follows.

On the other hand, since employees retire at the end of period 1, the 2-period contract cannot elicit work effort above $\bar{e}$ in period 1 since no further promotion is possible at that stage. To induce $e_1 > \bar{e}$ would require something like a retirement bonus paid to a specified proportion of retiring employees on retirement, and without that there would be some loss as compared with the sequence of 1-period contracts. In fact, of course, however many periods individuals work, it is only the final period to which this problem applies so that over a long working lifetime the difference becomes negligible. Since the reasons for the 2-period contract appear important in practice, that is the form of contract with payment based on an ordering of performance that will be considered formally in the rest of this paper. Moreover, since the complications of adding a third period of life in which individuals are retired (as would be required to model a retirement bonus properly) do not seem to add any essentially new insights, the possible use of such a bonus will be neglected.

In practice, employment contracts are not usually as fully specified as that suggested here. In an uncertain world there are, of course, costs to contracts that are inflexible unless they are complete in the sense of covering every eventuality. However, even with implicit or nonbinding agreements, there are clear advantages to dealing with terms verifiable by both parties. It is then immediately obvious when a firm is living up to its reputation or when customary procedures are being breached. This increases the cost to a firm of "cheating" in two ways. First, employee work effort will be adjusted straightaway, rather than with the lag that would occur if breaches were not immediately obvious. Second, employees can more quickly take other steps, such as collective action, to try to reimpose what they believe to be the implicit agreement. This latter aspect has been explored in Malcomson (1983) and hence will not be investigated further here. For the present, the terms of the 2-period contract are taken to be explicit and binding

Work Effort of Employees

Consider an individual offered the 2-period contract of the previous section with specified values of w_0, w_1, w_2, and P in a firm with a large number of employees. He will accept the contract if utility is at least $\bar{U}$. If he accepts the contract, he wor the wage w_0. If his observed effort in period 0 is gr defined by equation (1), he will be promoted and be in period 1; otherwise he will be offered wage w_1 in large number of other employees his own contrib work effort that, with P, determines x^* via (1) is determining his e_0 he will treat x^* as a parameter. (This is analogous to individual agents treating price as a parameter in a perfectly competitive market.)

An employee might not, however, wish to continue to work for the firm in period 1 if the wage he is offered is less than the wage other firms would be prepared to offer him for a 1-period contract in period 1 or less than his reservation wage for not working at all. Call the higher of these W_1. With a legally binding contract, the firm might be able to force an employee to continue to work, but it seems more realistic to assume either that it could not or that the costs of enforce-ing the contract are so high that it would not.[3] Define $w_1^* = \max\{w_1, W_1\}$. Then the employee will always determine period 0 behavior on the assumption that w_1^* is the wage he will receive in period 1 if not promoted. For reasons discussed earlier, $e_1 = \bar{e}$ always. Hence his expected utility, which will be denoted $V(e_0, w_0, w_1^*, w_2, x^*)$, is given by

$$\begin{aligned} V(e_0, w_0, w_1^*, w_2, x^*) &= F(x^* - e_0)U(e_0, w_0, \bar{e}, w_1^*) \\ &\quad + [1 - F(x^* - e_0)]U(e_0, w_0, \bar{e}, w_2). \end{aligned} \tag{2}$$

He will choose e_0 to maximize this subject to $e_0 \geqq \bar{e}$. For an interior solution for e_0, the first-order condition for this maximization is

$$\begin{aligned} f(x^* - e_0)[U(e_0, w_0, \bar{e}, w_2) &- U(e_0, w_0, \bar{e}, w_1^*)] \\ + F(x^* - e_0)&U_1(e_0, w_0, \bar{e}, w_1^*) \\ + [1 - F(x^* - e_0)]&U_1(e_0, w_0, \bar{e}, w_2) = 0. \end{aligned} \tag{3}$$

It is assumed that the utility function and density function have properties that ensure that the solution to (3) is unique. Denote that solution by $e^* = g(w_0, w_1^*, w_2, x^*)$. Then

$$e_0 = \max\{e^*, \bar{e}\}. \tag{4}$$

Note that, since $U_1(\cdot) < 0$ and $U_4(\cdot) > 0$, there can be an interior solution for e_0 only if $w_2 > w_1$ and if $f(x^* - e_0)$ is strictly positive so that there is a positive probability of promotion. This is as one would expect.

Crucial to the optimal contract that the firm offers is how effort responds to the variables in the contract. The contract does not, however, determine x^* directly. It specifies P, which, via (1), determines x^*. As argued above, for an individual acting competitively x^* is taken as given. But for a firm choosing the optimal contract, a change in P affects the work effort of all employees and hence, again via (1), x^*. The same applies to changes in w_0, w_1, and w_2. Hence the firm is interested in the effort supply function with P, not x^*, as an argument. Denote this effort supply function by

$$e_0 = h(w_0, w_1^*, w_2, P). \tag{5}$$

The derivatives of $h(\cdot)$ can be found from the first-order condition (3) using standard comparative static methods provided the effects of changes in effort on x^* are incorporated. Note from (1) that, for $0 < P < 1$, $\partial x^*/\partial e = 1$ and $\partial x^*/\partial P = -1/f(x^* - e)$. For $P = 0$ and $P = 1$, $f(x^*$

---

[3] In the present model, unlike that of Carmichael (1983), there is no reason for employees to switch firms because of mismatching, so there is no efficiency loss in tying an employee to a specific firm for 2 periods. The point here is concerned with the legal problem of preventing employees' switching for their own personal gain.

$- e) = 0$ and $\partial x^*/\partial P$ is infinite. Then

$$\left(V_{11} + V_{15}\frac{de_0}{de^*}\right)de^* + V_{12}dw_0 + V_{13}dw_1^* + V_{14}dw_2 - \frac{V_{15}}{f(x^* - e_0)}dP = 0, \tag{6}$$

where e_0, not e^*, enters the probability density function since it is the actual effort of employees that determines x^*. From this, assuming for simplicity that the utility function is additively separable between periods, one gets the following derivatives at an interior solution for e_0 and P:

$$h_1(\cdot) = \frac{-U_{12}}{U_{11}} \gtreqless 0 \text{ according as } U_{12} \gtreqless 0; \tag{7}$$

$$h_2(\cdot) = f(x^* - e_0)\frac{U_1}{U_{11}} < 0; \tag{8}$$

$$h_3(\cdot) = -f(x^* - e_0)\frac{U_1}{U_{11}} > 0; \tag{9}$$

$$\begin{aligned} h_4(\cdot) &= \frac{f'(x^* - e_0)[U(e_0, w_0, \bar{e}, w_2) - U(e_0, w_0, \bar{e}, w_1)]}{[f(x^* - e_0)U_{11}]} \\ &\gtreqless 0 \text{ according as } f'(x^* - e_0) \lesseqgtr 0 \text{ for } w_2 > w_1^*, \\ &= 0 \text{ for } w_2 = w_1^*, \end{aligned} \tag{10}$$

where U_1 must be evaluated at w_1^* for (8) and at w_2 f
as expected, increasing w_2 raises work effort while
ers it. The effect of w_0 depends on the sign of U
changing P is more complicated. The crucial eleme
probability density is increasing or decreasing w
For $f(\cdot)$ symmetric and unimodal, $h_4(\cdot) \gtreqless 0$ accordin
$P \lesseqgtr 1/2$).

The Optimal Contract for the Firm

Now consider the firm's choice of what contract to offer. It offers employees either 1-period contracts at fixed wages in both periods 0 and 1 of their working lives or a 2-period contract with a chance of promotion for period 1. In the latter case it could also hire period 1 employees on a 1-period contract at a fixed wage by offering sufficient to attract those who chose not to work, or who worked for other firms, in period 0.

Consider the 2-period contract denoted by $\{w_0, w_1, w_2, P\}$. To attract employees this must provide a minimum expected utility level $\bar{U}$, a constraint that can be expressed formally in terms of (2) with the effort supply function (5) inserted. For $w_1 \geqq W_1$, period 0 employees not promoted will return to work for the firm in period 1. For $w_1 < W_1$, they will not. Hence the firm's profit for each 2-period employee, $\pi(w_0, w_1, w_2, P)$, for any given 2-period contract is

$$\pi(w_0, w_1, w_2, P) = \begin{cases} h(w_0, w_1, w_2, P) - w_0 + r[s_1\bar{e} - Pw_2 \\ \quad - (1 - P)w_1], \quad \text{for } w_1 \geqq W_1; \\ h(w_0, W_1, w_2, P) - w_0 + rP[s_1\bar{e} - w_2], \\ \text{for } w_2 \geqq W_1 > w_1. \end{cases} \tag{11}$$

Clearly it could also be the case that $w_2 < W_1$, but this is equivalent to setting $P = 0$ in the lower part of (11) and hence does not need to be treated separately. The firm will choose its contract $\{w_0, w_1, w_2, P\}$ to maximize $\pi(\cdot)$ subject to the minimum expected utility constraint and the constraint $0 \leqq P \leqq 1$.

Now consider the alternative of hiring employees by 1-period, fixed-wage contracts only. In this case work effort is always $\bar{e}$ so, if the fixed wages for periods 0 and 1 are denoted w_0 and w_1, respectively, the firm's profit per employee is

$$\pi = \begin{cases} \bar{e} - w_0 + r(s_1\bar{e} - w_1), \quad \text{for } w_1 \geqq W_1; \\ \bar{e} - w_0, \quad \text{for } w_1 < W_1. \end{cases} \tag{12}$$

But for any given w_0 and w_1, (12) gives the same profit as (11) with $P = 0$ since $P = 0$ implies $h(\cdot) = \bar{e}$. Hence the firm could never be worse off by offering the 2-period contract, and therefore only the solution for the optimal 2-period contract needs to be considered here. Whenever that gives a profit per employee strictly greater for $0 < P < 1$ than for $P = 0$, the 2-period contract is strictly preferable to the 1-period, fixed-wage contracts.

The formal discussion of the optimal 2-period contract is relegated to the Appendix, and detailed proofs are presented there. Here the results of economic interest are summarized.

The first result (proposition 1 of the App.) is that for risk-loving employees the firm has strictly greater profits per employee from offering a contract with the proportion promoted, P, strictly between zero and one and $w_2 > w_1^*$. The intuition for this is obvious since, with risk-loving employees, offering a gamble enables the firm to reduce the cost of offering a given expected utility level. This is also optimal (proposition 2) for risk-neutral and risk-averse employees provided that the marginal utility of income in period 1 is sufficiently large relative to the rate at which the marginal disutility of effort in period 0 increases (i.e., provided $-U_4/U_{11}$ is sufficiently large) over the relevant range. Again the intuition is straightforward. From (9), for any given P with $0 < P < 1$, $-U_1/U_{11}$ determines the rate at which e_0 increases as w_2 increases, and if the increase in effort is sufficiently large it will more than compensate for the additional wage costs.

Since a 2-period contract with $0 < P < 1$ and $w_2 > w_1^*$ is a contract with payment based on an ordering of employee performance, these results establish that, under the conditions of propositions 1 and 2, such a contract dominates contracts based on individual performance only, of which the only examples available in the current context are the 1-period, fixed-wage contracts or the 2-period contract with $P = 0$, $P = 1$, or $w_2 = w_1^*$. Moreover, with $0 < P < 1$ and $w_2 > w_1^*$, there is genuine internal promotion in the model since only employees working for the firm in period 0 will be considered for the high wage w_2 in period 1. New employees enter the firm only at w_0 (or at their marginal product $s_1\bar{e}$ if they worked for another firm in period 0). Hence, the model generates the characteristics of an internal labor market, feature 2 from the Introduction. Indeed, in this model the only promotion is internal promotion, but that obviously may not be the case in more general models.

To see how the model ties up with feature 3 from the Introduction, consider the zero-profit condition that must hold in equilibrium. This is given by setting $\pi(\cdot) = 0$ in equation (11). Note that employees are free to work for a different firm in period 1 from their employer of period 0 if they can get a higher wage. Moreover, other firms would always be prepared to offer them a wage equal to their marginal product in that period so that, as long as there are other firms [illegible] in which they are equally productive, they would be [illegible] that is, $W_1 \geqq s_1\bar{e}$. With $w_2 > w_1$, it follows immediately [illegible] brackets in (11) must be negative. But in both upper [illegible] of (11) this term is the difference between margin[illegible] average wage in period 1 for employees on the 2[illegible] Hence 2-period employees are, on average, paid [illegible] marginal product in period 1. This is, as can be seen by [illegible] 0 in (11), exactly offset in present value by their being paid [illegible] their marginal product in period 0 so that, over their working life, the present value of expected wages equals the present value of marginal product. Hence the ratio of the average wage to marginal product rises over employees' working lives, which is precisely the relationship described under feature 1. Whether the real wage itself rises or falls between periods 0 and 1 depends on a number of factors. Individual effort falls since $e_0 \geqq \bar{e}$, but this is countered by any experience-related increases in skill, that is, $s_1 > 1$. Medoff and Abraham (1980) find no significant relationship between experience and productivity within a given grade, and for the present model to be consistent with that, the experience-related increase in skill would have to just counterbalance the decline in work effort, that is, $e_0 = s_1\bar{e}$. It then follows from what was said above that $w_2 > w_1 > w_0$, and hence wages rise with length of service. Of course, since e_0 is endogenous to the model, one can express the condition $e_0 = s_1\bar{e}$ in terms of the parameters, but there seems little to be learned from that. Note also that the variance of earnings increases with experience (feature 4) since all period 0 employees earn w_0 whereas some period 1 employees earn w_2 and some w_1^*.

In view of the relationship between the wage and the marginal product it generates, the model preserves Lazear's (1979) insight into the reasons for mandatory retirement. Formally this can be included in a straightforward way by incorporating a third period into the model in which an individual's reservation wage is always greater than his productivity. Then, as in Lazear's model, at the time the employment contract is made both parties would agree to the contract specifying retirement at the end of the second period since this gives the maximum gain to be shared between the parties. But that does not mean that individuals would choose to retire at the end of the second period if they were able to continue at their second period wage since that wage is above their marginal product and hence may be above their reservation wage. This is the essence of Lazear's point.

It is also interesting to note that, since period 2 employees are paid more than their marginal product, the firm would always have a short-term interest in laying them off, although it would obviously undermine the credibility of the incentive structure if it did so. This makes it clear why the firm would be prepared to bind itself to last-in-first-out or seniority rules for determining the order of layoffs and redundancies.

Next consider the question of hierarchy. At any given date there are two generations of individuals employed by the firm. All period 0 employees are paid w_0 whereas period 1 employees are divided between those earning w_1 and those earning w_2. As long as the firm is not declining in size, it is certainly true that there are more employees earning w_0 than w_1 or w_2. With $w_2 > w_1 > w_0$ there are then certainly more employees at the bottom of the hierarchy than in either of the upper echelons. To have fewer at the top than in the middle, however, requires $P < 1/2$, so that there are fewer earning w_2 than w_1. Conditions under which this happens are explored in the Appendix. Unfortunately, I have been unable to find a general way to characterize the economic conditions under which it occurs. Certainly for $f(\cdot)$ a normal distribution, $P < 1/2$ for a not implausible class of utility functions. One example is when the utility function for income in period 1 displays constant relative risk aversion between 0 and $1/2$. See the Appendix for details.

This shows how the present model can generate the first four features of employment structures discussed in the Introduction. The fifth is slightly more tricky. The sense in which it fits easily is the following. To enhance the distinction between those promoted and those not promoted, and hence simplify the employees' monitoring of firm behavior, the firm would wish to label them as two different grades, to which the wages w_1 and w_2 are attached. Then the wage an individual receives depends only on what grade he is in since in all other respects the individuals in the two grades are identical. There is obviously no scope for individual bargaining over the rate for his grade, since wages in each period do not depend directly on productivity in that period but, at least for period 1, on the incentive effects the firm wishes to generate for the previous period. Moreover, the wage for any given grade is not directly related to external market wages, though obviously there is an indirect connection through the level of utility $\bar{U}$ that a firm's contract must offer. The relationships among w_0, w_1, and w_2 are determined by the incentive requirements of the firm. This would appear consonant with differentials being determined by administrative procedures, as discussed in Doeringer and Piore (1971), rather than by external market forces.

It is not quite the same, however, as wages being attached to jobs since, strictly speaking, within the present model everybody is doing the same job. That, of course, reflects the simplicity of the model. If, for example, one introduced supervisors of the type considered by Calvo and Wellisz (1979), who would have an obvious role to play monitoring $\bar{e}$, it may well make sense for the firm to use those promoted in a supervisory capacity as a means of reinforcing the distinction between those promoted and those not promoted. [illegible] wage paid would be attached not just to the grade but [illegible] which an individual is assigned.

Concluding Remarks

It has been shown that the use of contracts with pay[illegible] ranking of employees' performance can be made enfor[illegible] means of providing performance incentives even in a situation of asymmetric information in which employees cannot verify an employer's observations of their performance so that a contract specifying a wage directly dependent on individual performance is unenforceable. Under certain conditions 2-period contracts with promotion are preferable to 1-period and 2-period fixed-wage contracts, which are the only systems with payment by standard enforceable in the model. Note that this result is in no way dependent on the existence of stochastic elements in performance correlated across employees and so provides a wider set of circumstances than those discussed by Green and Stokey (1983) under which contracts based on an ordering of employees' performance are preferred to contracts based on individual performance only. Moreover, it has been shown how the internal incentive structure of such contracts can generate the five features of labor organization outlined in the Introduction.

It is also of interest to consider the properties of a competitive equilibrium in a model with 2-period contracts of the sort analyzed here. When the conditions for propositions 1 and 2 in the Appendix hold, such an equilibrium is a strict Pareto improvement over a competitive equilibrium with payment based on individual performance only. The reason is clear. In a competitive equilibrium with constant returns to scale, firms always earn zero profits per employee so that they are never worse off under 2-period contracts with promotion. But when such contracts give higher profits per employee for a given level of employees' expected utility, an equilibrium with zero profits gives higher expected utility for employees. Hence employees are better off.

One final point to note is that it is straightforward to generalize the model used here to many different types of individuals. Indeed, if one indexes the utility functions, effort levels, and contracts by i, all the arguments used above can be repeated for each i. It is, however, interesting to note that the different types of individuals would in general be offered different 2-period contracts with different profiles of lifetime earnings and probabilities of promotion so that, in practice, one would expect a variety of different job ladder structures to emerge.

Appendix

This Appendix is concerned with the nature of the optimal contract arising from the choice of w_0, w_1, w_2, and P to maximize the firm's profit per employee, $\pi(\cdot)$ defined by equation (11), subject to the minimum expected utility constraint and the constraint $0 \leqq P \leqq 1$. Note that, given additive separability between periods and $e_1 = \bar{e}$ exogenous, the utility function may be written

$$U(e_0, y_0, \bar{e}, y_1) = U^1(e_0, y_0) + U^2(y_1).$$

Then the minimum expected utility constraint is given by

$$U^1[h(w_0, w_1^*, w_2, P), w_0] + (1 - P)U^2(w_1^*) + PU^2(w_2) \geqq \bar{U}. \qquad \text{(A1)}$$

Analyzing this problem is somewhat complicated by the fact that the expression for profit given in (11) is not differentiable at $w_1 = W_1$.

The initial concern is to show under what conditions the optimal contract has $0 < P < 1$ and $w_2 > w_1^*$. It is useful to define the class of contracts in which employees bear no risk.

DEFINITION: A certainty contract is one with $P = 0$, $P = 1$, or $w_2 = w_1^*$.

Since for any certainty contract with $w_2 = w_1^*$ or $P = 1$ there is always an equivalent certainty contract with $P = 0$, only the latter case need be of concern here. As argued in the text, for a certainty contract $e_0 = \bar{e}$ always.

Proposition 1: For risk-loving employees, the optimal contract has $0 < P < 1$ and $w_2 > w_1^*$.

Proof: Any certainty contract can always be written in a form with $P = 0$ and $w_2 > w_1^*$. Hence it suffices to rule out that this is optimal. For $P = 0$, $e_0 = \bar{e}$ always. For positive P and $w_2 > w_1^*$, work effort is certainly no less than $\bar{e}$. But for risk-loving employees, the actuarial cost to the firm of providing employees with expected utility $\bar{U}$ is less if the wage for period 1 involves a gamble than if it is certain, and this can always be achieved by raising P and lowering w_0. Hence there exists a $P > 0$ that gives at least as much output and lower cost than $P = 0$.

Remark: The obvious route to proving this result, that is, raising w_2 and lowering w_1 to leave the actuarial cost of wages to the firm unchanged, is not always available. The reason is that it may be that $w_1 = W_1 = s_1\bar{e}$ for a certainty contract so that lowering w_1 would merely ensure that employees not promoted left the firm and then the firm's profit per employee would necessarily fall.

For risk-averse employees the equivalent result is more complicated The reason is that then the firm's profit per employee always has a local maximum at that certainty contract that, over the class of certainty contracts, maximizes the firm's profits per employee subject to the expected utility constraint This is because under a certainty contract the constraint $e_0 \geqq \bar{e}$ is always strictly binding; that is, the solution to (3) always has $e^* < \bar{e}$, so that a local variation in the contract never increases work effort and any change in the contract that satisfies the expected utility constraint by introducing a gamble always raises the actuarial cost of wages to the firm. Hence one needs to consider
on profits per employee of a discrete change in the cont
the assumptions in the main text, it is assumed that
bounded away from zero. The following lemma is used

Lemma 1: For some $0 < P < 1$, there exists a w_2 such that
the solution to (3) is $e^* = \bar{e}$ for any finite $\bar{e}$

Proof: For $e^* < \bar{e}$, it follows from (6) that

$$\frac{\partial e^*}{\partial w_2} = -\frac{V_{14}}{V_{11}} = -\frac{f(\cdot)U^{2\prime}(w_2)}{U_{11} - f'(\cdot)[U^2(w_2) - U}$$

For given P, $f(\cdot)$ and $f'(\cdot)$ are fixed, and there certainly exists a $0 < P < 1$ such that $f(\cdot) > 0$ and $f'(\cdot) > 0$. Then, for $w_2 > w_1^*$, $\partial e^*/\partial w_2 > 0$ and is bounded away from zero. Hence there exists a w_2 such that e^* exceeds any given number $\bar{e}$, so proving the lemma

Proposition 2: For risk-neutral and risk-averse employees, there exist utility functions such that the optimal contract has $0 < P < 1$ and $w_2 > w_1^*$ for finite $\bar{e}$.

Proof: Suppose the optimal contract were a certainty contract Then, since any certainty contract has an equivalent certainty contract with $w_2 = w_1^*$ and P arbitrary, without loss of generality denote the optimal contract by $\{\bar{w}_0, \bar{w}_1, \bar{w}_1, P\}$ with P arbitrary. Choose a $\bar{P}$ for which lemma 1 holds and let $w_2^*(\bar{w}_0, \bar{w}_1, \bar{P})$ be the w_2 defined in lemma 1 for which $e^* = \bar{e}$. Now consider an increase in w_2 from $\bar{w}_1$. This cannot violate the minimum expected utility constraint, since employees can always choose effort $\bar{e}$ and be certain of income $\bar{w}_1$ in period 1. For any $w_2 < w_2^*$, profit per employee π falls since there is no increase in work effort. For any $w_2 > w_2^*$,

$$\Delta\pi = \int_{w_2^*}^{w_2} \left(\frac{\partial e_0}{\partial w_2}\right) dw_2 - rP(w_2 - \bar{w}_1).$$

But it follows from (6) that, for $w_2 > w_2^*$,

$$\frac{\partial e_0}{\partial w_2} = -\frac{f(x^* - e_0)U^{2\prime}(w_2)}{U_{11}(\cdot)},$$

so that since, for given P, $x^* - e$ is given,

$$\Delta\pi = f(x^* - e_0) \int_{w_2^*}^{w_2} \left[\frac{U^{2\prime}(w_2)}{-U_{11}(\cdot)}\right] dw_2 - rP(w_2 - \bar{w}_1).$$

Since there is no restriction on the size of $U^{2\prime}(w_2)/-U_{11}(\cdot)$, there certainly exist utility functions with sufficiently large values of this that $\Delta\pi > 0$, contradicting that the certainty contract was optimal.

REMARK: Formally, imposing the condition that the utility function satisfies

$$\max_{w_2, P} \; f(x^* - e_0) \int_{w_2^*(\bar{w}_0, \bar{w}_1, P)}^{w_2} \left[\frac{U^{2\prime}(w_2)}{-U_{11}(\cdot)}\right] dw_2 - rP(w_2 - \bar{w}_1) > 0$$

is sufficient for the proof of the proposition.

The remaining proposition to be proved concerns conditions under which $P < 1/2$. For any certainty contract this is trivial since there is always a precisely equivalent contract with $P = 0$ that certainly satisfies $P < 1/2$. Consider therefore the other cases. The firm's profit per employee, $\pi(\cdot)$ given by (11), is a differentiable function of the contract variables except at $w_1 = W_1$. Let λ denote the Lagrange multiplier attached to the minimum expected utility constraint (A1). Then, for any contract maximizing (11) subject to (A1) with $0 < P < 1$ and $w_2 > w_1^*$, the following first-order necessary conditions must be satisfied, with inequalities bracketed together being complementary:

$$[1 + \lambda U_1^1(\cdot)]h_1(\cdot) - 1 + \lambda U_2^1(\cdot) = 0; \qquad \text{(A2)}$$

$$\left.\begin{aligned}[1 + \lambda U_1^1(\cdot)]h_2(\cdot) + (1 - P)[\lambda U^{2\prime}(w_1) - r] &\leqq 0\\ w_1 &\geqq W_1\end{aligned}\right\} \text{ if } w_1 \geqq W_1; \qquad \text{(A3)}$$

$$[1 + \lambda U_1^1(\cdot)]h_3(\cdot) + P[\lambda U^{2\prime}(w_2) - r] = 0; \qquad \text{(A4)}$$

$$[1 + \lambda U_1^1(\cdot)]h_4(\cdot) + \{\lambda[U^2(w_2) - U^2(w_1)] - r(w_2 - w_1)\} = 0; \qquad \text{(A5)}$$

$$\left.\begin{aligned}U^1[h(\cdot), w_0] + (1 - P)U^2(w_1) + PU^2(w_2) &\geqq \bar{U}\\ \lambda &\geqq 0\end{aligned}\right\}. \qquad \text{(A6)}$$

Note from (11) that in the case $W_1 = s_1\bar{e}$ (i.e., employees would choose to work in period 1 at a wage equal to their marginal product), the firm would always have lower profits for $w_1 < W_1$ than for $w_1 = W_1$.

Before deriving the remaining proposition it is useful to establish the following lemmas:

LEMMA 2: For risk-averse employees, an optimal contract with $w_2 > w_1 \geqq W_1$ and $0 < P < 1$ implies

$$\lambda U^{2\prime}(w_1) - r > 0, \; \lambda U^{2\prime}(w_2) - r < 0, \text{ and } 1 + \lambda U_1^1(\cdot) > 0. \qquad \text{(A7)}$$

Proof: From (8) and (9) of the text, $h_2(\cdot) < 0$ and $h_3(\cdot) > 0$. For a risk averter, $U^{2'}(w_1) > U^{2'}(w_2)$. Hence (A3) and (A4) can both be satisfied only if (A7) holds.

Lemma 3: For risk-averse employees, an optimal contract with $w_2 > w_1 \geqq W_1$ implies that $\lambda > 0$ and hence that the minimum expected utility constraint is binding.

Proof: Follows directly from lemma 2 and the fact that $U^{2'}(\cdot)$, $r > 0$.

Lemma 4: For an optimal contract with $w_1 \geqq W_1$,

$$\lambda \leqq r\left[\frac{1-P}{U^{2'}(w_1)} + \frac{P}{U^{2'}(w_2)}\right]$$

with equality holding when $w_1 > W_1$.

Proof: Divide (A3) by $h_2(\cdot)$, noting that $h_2(\cdot) < 0$ so that the inequality is reversed, and (A4) by $h_3(\cdot)$. Subtract the latter expression from the former to give

$$\frac{(1-P)[\lambda U^{2'}(w_1) - r]}{h_2(\cdot)} - \frac{P[\lambda U^{2'}(w_2) - r]}{h_3(\cdot)} \geqq 0$$

or

$$\lambda\left[\frac{(1-P)U^{2'}(w_1)}{h_2(\cdot)} - \frac{PU^{2'}(w_2)}{h_3(\cdot)}\right] \geqq r\left[\frac{1-P}{h_2(\cdot)} - \frac{P}{h_3(\cdot)}\right].$$

Substitution of (8) and (9) of the text for $h_2(\cdot)$ and $h_3(\cdot)$ with appropriate cancellation of terms—note that $U^{2'}(w) \equiv U'_4(\cdot)$—yields the required result.

Lemma 5: For an optimal contract with $w_1 > W_1$,

$$\frac{\partial\lambda}{\partial P} = r\left[\frac{1}{U^{2'}(w_2)} - \frac{1}{U^{2'}(w_1)}\right] \gtreqless 0$$

according as employees are risk averse, risk neutral, or risk loving

Proof. Follows from lemma 4 and definitions of risk aver[illegible] trality, and risk loving.

Proposition 3: For $W_1 = s_1\bar{e}$, $f(x - e_0)$ normal, U_{111} [illegible] displaying risk aversion and satisfying

$$\tfrac{1}{2}\left[\frac{1}{U^{2'}(w_1)} + \frac{1}{U^{2'}(w_2)}\right][U^{2}(w_2) - U^{2}(w_1)] - (w_2 - w_1$$

the optimal contract has $P < \frac{1}{2}$.

Proof: For any certainty contract there is always an equivalent contract with $P = 0$, so for certainty contracts the proposition is trivially true. For $W_1 = s_1\bar{e}$, $w_1 \geqq W_1$ always, as argued above, so that for any contract other than a certainty contract, the conditions for lemmas 2–4 are satisfied. From (10) of the text and $f(x - e_0)$ normal, it follows that $h_4(\cdot) \gtreqless 0$ according as $P \lesseqgtr \frac{1}{2}$ and also that, when $U_{111}(\cdot) \leqq 0$, $h_{41}(\cdot) < 0$. By lemma 2, therefore,

$$X_1 \equiv [1 + \lambda U^1_1(\cdot)]h_4(\cdot) \lesseqgtr 0 \text{ according as } P \gtreqless \tfrac{1}{2}, \qquad \text{(A9)}$$

where X_1 is defined implicitly. Since for any given w_1 and w_2 $(w_2 > w_1)$ and, in particular, an optimal w_1 and w_2, $h_4(\cdot) \to \infty$ as $P \to 0$, it follows that $X_1 \to \infty$ as $P \to 0$. Hence, since the optimal contract has $w_2 > w_1$, X_1 can take on any positive value at the optimal w_1 and w_2 for some $0 < P < \frac{1}{2}$. But, when (A8) holds and for any given $w_2 > w_1$, it follows from lemma 4 and $r > 0$ that

$$X_2 \equiv \lambda[U^2(w_2) - U^2(w_1)] - r(w_2 - w_1) < 0 \text{ for } P = \tfrac{1}{2}, \qquad \text{(A10)}$$

where X_2 is implicitly defined, and from lemma 5 that, when $w_1 > W_1$, $\partial X_2/\partial P > 0$. Hence, at an optimal w_0, w_1, and w_2, $X_2 < 0$ for any $0 < P < \frac{1}{2}$. Equation (A5) requires that, at an optimum with $0 < P < 1, X_1 = -X_2$, which must be the case for some $0 < P < \frac{1}{2}$. But $h_{44}(\cdot) < 0$ ensures that there is only one P for which this is true at any given w_0, w_1, and w_2. Hence for $w_1 > W_1$ the proposition holds. For $w_1 = W_1$, it follows from lemma 4 that, for any given P, w_1, and w_2, X_2 is bounded above by its value for $w_1 > W_1$, so the conclusion holds for that case too.

REMARK: The conditions for proposition 3 are only sufficient conditions. In fact $f(x^* - e_0)$ normal and $U_{111}(\cdot) \leqq 0$ are much stronger than required but have the virtue of easy interpretability. Unfortunately I have found no economic interpretation of (A8). It is, however, satisfied by certain plausible classes of utility functions. One example is the class of constant relative risk aversion utility functions $U^2(w) = w^a$ for $\frac{1}{2} < a < 1$.

References

Alchian, Armen A., and Demsetz, Harold. "Production, Information Costs, and Economic Organization." *A.E.R.* 62 (December 1972): 777–95.

Calvo, Guillermo A., and Wellisz, Stanislaw. "Hierarchy, Ability, and Income Distribution." *J.P.E.* 87, no. 5, pt. 1 (October 1979): 991–1010.

Carmichael, Lorne. "Firm-specific Human Capital and Promotion Ladders." *Bell J. Econ.* 14 (Spring 1983): 251–58.

Doeringer, Peter B., and Piore, Michael J. *Internal Labor Markets and Manpower Analysis.* Lexington, Mass.: Lexington Books, 1971.

Green, Jerry R., and Stokey, Nancy L. "A Comparison of Tournaments and Contracts." *J.P.E.* 91 (June 1983): 349–64.

Greenwald, Bruce C. N. *Adverse Selection in the Labor Market.* New York: Garland, 1979.

Grossman, Sanford J., and Hart, Oliver D. "An Analysis of the Principal-Agent Problem." *Econometrica* 51 (January 1983): 7–45.

Harris, Milton, and Holmstrom, Bengt. "A Theory of Wage Dynamics." *Rev Econ. Studies* 49 (July 1982): 315–33.

Holmstrom, Bengt. "Equilibrium Long-Term Labor Contracts." Discussion Paper no. 414. Evanston, Ill.: Northwestern Univ., Kellogg Grad. School of Management, January 1981.

Lazear, Edward P. "Why Is There Mandatory Retirement?" *J.P.E.* 87 (December 1979): 1261–84.

———. "Agency, Earnings Profiles, Productivity, and Hours Restrictions." *A.E.R.* 71 (September 1981): 606–20.

Lazear, Edward P., and Rosen, Sherwin. "Rank-Order Tournaments as Optimum Labor Contracts." *J.P.E.* 89 (October 1981): 841–64.

Lydall, Harold F. *The Structure of Earnings.* Oxford: Clarendon, 1968.

Malcomson, James M. "Trade Unions and Economic Efficiency." *Econ. J.* 93 (suppl.; March 1983): 51–65.

Medoff, James L., and Abraham, Katharine G. "Experience, Performance, and Earnings." *Q.J.E.* 95 (December 1980): 703–36.

Mincer, Jacob A. *Schooling, Experience, and Earnings*. New York: Columbia Univ. Press (for N.B.E.R.), 1974.

Radner, Roy. "Monitoring Cooperative Agreements in a Repeated Principal-Agent Relationship." *Econometrica* 49 (September 1981): 1127–48.

Simon, Herbert A. "The Compensation of Executives." *Sociometry* 20 (March 1957). 32–35.

Stiglitz, Joseph E. "Incentives, Risk and Information: Notes towards a Theory of Hierarchy." *Bell J. Econ.* 6 (Autumn 1975): 552–79.

Williamson, Oliver E.; Wachter, Michael L.; and Harris, Jeffrey E. "Understanding the Employment Relation: The Analysis of Idiosyncratic Exchange." *Bell J. Econ.* 6 (Spring 1975): 250–78

Sraffa, Labor Theories of Value, and the Economics of Real Wage Rate Determination

Edwin Burmeister

University of Virginia

The labor theory of value and the question of whether or not a real wage rate versus profit rate trade-off exists have been topics debated since the times of Ricardo and Marx. More recently these topics have been investigated using Sraffa's standard commodity approach. It is argued here that such approaches are basically flawed. However, it also is shown that there are situations consistent with perfect competition for which negatively sloped real wage rate versus profit or interest rate trade-offs do exist. Thus, despite perfect competition, genuine bargaining situations may arise between workers and owners of the capitalistic means of production.

I. Introduction

It is well known that the classical labor theory of value holds in its purest form (i.e., that at every profit rate the ratio of the prices of any two commodities is equal to the ratio of their total embodied labor costs) if and only if the capital structure of the economy satisfies Marx's condition for equal organic composition of capital.[1] Under

---

An earlier version of this paper was presented at the History of Economics Society Convention, May 23–26, 1982, Duke University, Durham, North Carolina, and some passages are taken from a series of lectures I gave at the University of Chicago in 1980. I am grateful for comments and advice received from Gary Becker, H. Gregg Lewis, Robert E. Lucas, Jr., Marjorie B. McElroy, Ron Michener, and especially Paul A. Samuelson and John Whitaker. Likewise, suggestions from a referee have resulted in substantial improvements. This research was supported by NSF research grant SES-8218229.

[1] A superb reference is Samuelson (1972), where the labor theory of value is explained with simple numerical examples. Other classic references include Stigler (1958) and Ricardo (1963).

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0002$01.50

these stringent conditions, the real wage measured in terms of any commodity as numeraire is linearly and inversely related to the rate of profit. One of Sraffa's (1960) contributions was to demonstrate the existence of a standard commodity such that if a "real" wage is measured in terms of this standard commodity, then the linear "real" wage–profit trade-off is preserved. Accordingly, Sraffa enables one to construct a different labor theory of value, upon which various notions of exploitation can be based, without the stringent equal organic composition of capital restriction.

However, there are several economic objections to Sraffa's standard commodity approach, and I have selected the following issues for discussion:[2]

1. Sraffa's weights used to construct his basket of goods are seen to be determined completely from the technology without regard for consumer preferences.
2. A unique standard commodity does not exist when there is more than one production technique available.
3. No primary factors other than labor can be admitted.
4. Even when there is only one technique of production and labor is the only primary factor, generally Sraffa's standard commodity does not exist when there are durable capital goods; only a subset of the possible depreciation patterns will admit a standard commodity.
5. Comparisons of alternative steady-state equilibria—the basis of most labor theories of value as well as Sraffa's work—are fictitious exercises because they do not reflect the feasible intertemporal choices open to an economy.

What, then, are we to make in general of the propo
exists an inverse trade-off between real wage and r
that workers can be exploited by forcing a lower rea
higher interest rate? A primary contribution of this
how such trade-offs can exist in a perfectly competi
technology is piecewise differentiable, the margin
temporal consumption transformation between any two
provide technologically determined bounds on the real inter
consistent with perfect competition. Then it is shown that under extremely general conditions an inverse relationship between the real

---

[2] It is recognized that Sraffa was not writing in the neoclassical tradition and that he was aware that some, or perhaps even all, of the conditions listed below caused complications. Whether or not these complications severely limit the economic usefulness of Sraffa's approach, or whether the objections discussed here can be dismissed, is a matter left for the reader to decide. Those interested in recent views more sympathetic to labor theories of value are referred to Morishima (1973), Fatwell (1975), Nell (1980), Harcourt (1982), Parys (1982), and Flaschel (1983), though this list is not meant to be exhaustive.

wage and interest rates exists between these bounds. Moreover, results are derived both for any intertemporally competitive path and for the special case of a steady-state path. The conclusion is robust with respect to assumptions about joint production, heterogeneous capital goods, and depreciation patterns and therefore represents a fundamental proposition about competitive economies.

Real wage rate versus interest rate trade-offs do not exist if the technology is differentiable everywhere, and the degree of indeterminacy consistent with a competitive path shrinks as the technology becomes "smoother." But in instances where there are legitimate real wage–interest rate trade-offs for competitive models, owners of the means of production may "exploit" workers by offering only the lowest wage consistent with the binding inequalities. Likewise, unions may exercise "power" by refusing to supply labor at wages lower than the highest value consistent with those binding inequalities. Some type of game-theory solution is required to resolve this indeterminacy.

II. Equal Organic Composition of Capital

As a preliminary I shall introduce some familiar notation and review the standard technological assumptions implicit in Sraffa's model.[3] Consider an economy in which n commodities are produced by themselves and labor under constant returns to scale by n nonjoint production functions. In the simplest case these production functions define a single Leontief-Sraffa production technique

$$\left(\frac{a_0}{a}\right) = \left[\begin{array}{ccc} a_{01} & \cdots & a_{0n} \\ \hdashline a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{array}\right],$$

where the a_{ij}'s are fixed input coefficients indicating the quantity of factor i required to produce one unit of commodity j and where the subscript zero denotes a labor input. It is well known that total embodied labor inputs are given by the vector

$$a_0 + a_0 a + a_0 a^2 + \ldots = a_0(I - a)^{-1} \equiv (l_1, \ldots, l_n) = l; \quad (1)$$

see, for example, Dorfman, Samuelson, and Solow (1958, pp. 253–

---

[3] For expositional clarity I shall avoid considering the complications caused by joint production, except below in Sec. IV where I briefly discuss the problems that arise when older machines are produced jointly with final output. Interested readers are referred to Samuelson (1983, 1984), Baldone (1984), and Burmeister and Samuelson (1984)

54), Bruno, Burmeister, and Sheshinski (1966, pp. 544–45), and Samuelson and von Weizsäcker (1972). It is also well known that these total embodied labor requirements are equal to prices at a zero interest or profit rate. To see this fact, write the price = cost condition as

$$P_j = a_{0j}W + (1 + r)\sum_{i=1}^{n} P_i a_{ij}, \quad j = 1, \ldots, n, \tag{2}$$

where P_j is the price of commodity j, W is the nominal post factum wage rate, and r is the interest or profit rate, and where for simplicity we follow the first part of Sraffa and postulate a technology with circulating capital. Defining the price vector $p = (p_1, \ldots, p_n) = (P_1/W, \ldots, P_n/W)$, we write (2) as

$$p = a_0 + (1 + r)pa, \tag{3}$$

which has a solution

$$p(r) = a_0[I - (1 + r)a]^{-1}, \quad 0 \leq r < r^*. \tag{4}$$

The assumption that r^* is positive is equivalent to the condition that the economy is productive and capable of generating positive steady-state equilibrium p_i's at a zero rate of profit. Substituting $r = 0$ in (4) gives the total embodied labor requirements (1), that is,

$$p(0) = l. \tag{5}$$

For positive profit rates, steady-state equilibrium prices given by (4) do not equal total embodied labor requirements, that is,

$$p(r) \neq l, \quad r > 0.$$

However, it still may be true that relative prices [illegible] embodied requirements in the sense that

$$\frac{p_i(r)}{p_j(r)} = \frac{l_i}{l_j}, \quad 0 \leq r < r^*.$$

To put the matter another way, if relative prices are [illegible] the profit rate, then they must equal relative prices at the [illegible] profit rate $r = 0$, and hence they must equal relative total embodied labor requirements; that is, if relative prices are independent of r, we have

$$\frac{p_i(r)}{p_j(r)} = \frac{p_i(0)}{p_j(0)} = \frac{l_i}{l_j}. \tag{8}$$

Conversely, the pure labor theory of value expressed by (7) or (8) cannot hold whenever relative prices change with the profit rate.

Thus if and only if relative prices do not change with the interest or profit rate r, price ratios are equal to the ratios of total embodied labor inputs, as asserted by equation (8) above.

It is well known that the necessary and sufficient conditions for this labor theory of value to hold are exceedingly stringent. They may be summarized in the following form.

EQUAL ORGANIC COMPOSITION OF CAPITAL THEOREM: Relative prices in any units of measurement are constant for all feasible r if and only if there is equal organic composition of capital; that is, each "labor/value of capital" ratio is the same function of r for every industry with

$$\frac{WL_j}{\sum_{i=1}^{n} P_i(r)K_{ij}} = \frac{a_{0j}}{\sum_{i=1}^{n} P_i(r)a_{ij}} = r^* - r, \quad j = 1, \dots, n. \tag{9}$$

Here $r^* = (1/\lambda^*) - 1$ and λ^* is the Frobenius root of the matrix a, L_j is the direct labor used to produce commodity j, and K_{ij} is the amount of type i capital used to produce commodity j.

One of the many proofs of this result can be found in Burmeister (1980*c*); it is, of course, originally due to Marx.

Now let $\omega_i(r)$ denote the real wage rate in terms of the ith commodity as numeraire; that is, define $\omega_i(r) \equiv W/P_i(r)$, for $i = 1, \dots, n$. The important economic consequence of equal organic composition of capital is that then every factor-price frontier, which gives the tradeoff between $\omega_i(r)$ and r, is linear:

$$\omega_i(r) = \frac{1}{p_i(r)} = \frac{a_{0i}[1 - (r/r^*)]}{(1 + r^*)/r^*}, \quad i = 1, \dots, n. \tag{10}$$

These real wage rates decrease from their maxima $\omega_i(0) = a_{0i}r^*/(1 + r^*)$, for $i = 1, \dots, n$, at a zero profit or interest or profit rate ($r = 0$) to their minima $\omega_i(r^*) = 0$ at the maximal interest or profit rate ($r = r^*$), as illustrated in figure 1.

Now suppose we define a new real wage in terms of *any* feasible consumption basket $(c_1, \dots, c_n)$ by

$$w \equiv \frac{W}{\sum_{i=1}^{n} P_i c_i} = \frac{1}{\sum_{i=1}^{n} p_i c_i}.$$

It then can be shown that for any consumption basket weights, this real wage is also linearly related to the interest or profit rate with

$$r = r^*(1 - w), \tag{11}$$

provided the labor theory of value is valid (see, e.g., Burmeister 1968). Equation (11) is illustrated in figure 2.

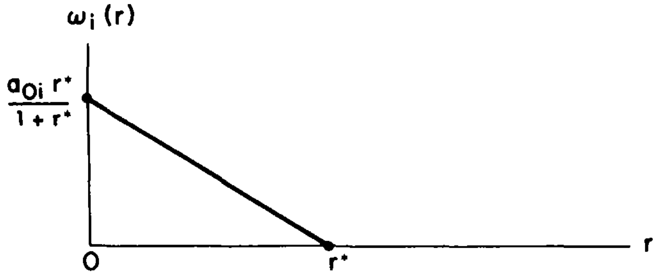


FIG. 1.—The real wage rate ω_i is a linear function of the profit or interest rate r when the labor theory of value is valid.

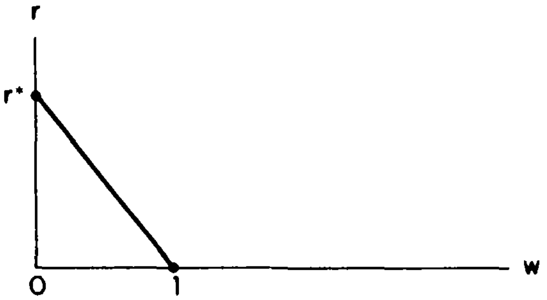


FIG. 2.—Linear relationship between the profit rate r and the real wage rate w for any consumption basket weights when the labor theory of value is valid.

Many economists have recognized that in realistic situat
prices do differ from true labor-value ratios (e.g., t
[1958] paper), but there is little empirical work c
with the extent of such discrepancies. However, t
dence implied, for example, by the works of Ripley
Sellekaerts and Lesage (1973), and the survey
strongly suggests huge errors.[1]

It is also interesting to note that equal organic compositi
tal is one sufficient condition for aggregation of heterogeneous capital goods into a well-behaved index of "capital" (see Brown and Chang 1976; Burmeister 1976, 1980*b*).

---

[1] *A condition equivalent to equal organic composition of capital is that labor's relative* share (WL_i/P_iQ_i, where Q_i denotes the total output of the ith commodity) be the *same* function of the interest rate for each $i = 1, \ldots, n$ (see Burmeister 1980*c*, p. 95). Thus empirical cross-section data with the interest rate constant and widely differing labor shares are sufficient to refute the labor theory of value, and a glance at national income data provides convincing evidence. An alternative simple test for equal organic composition of capital is given by Parys (1977).

III. Sraffa's Standard Commodity

Sraffa's celebrated contribution was to invent a different "labor theory of value" by constructing a new wage measure such that the linearity depicted in figure 2 remains valid even when there is not equal organic composition of capital. In particular, Sraffa found particular consumption basket weights $(c_1^*, \ldots, c_n^*)$ to define what I shall call "Sraffa's wage," denoted by w^s. This new wage rate measure is given by

$$w^s \equiv \frac{1}{\sum_{i=1}^{n} p_i c_i^*}. \tag{12}$$

The remarkable result proved by Sraffa (1960) is that

$$r = r^*(1 - w^s) \tag{13}$$

holds even when there is not equal organic composition of capital and the ordinary labor theory of value is false. Accordingly, all one needs to do is replace w by w^s in figure 2, and all the Marxian analysis that is based on the linearity of figure 2 can proceed free of the objectionable equal organic composition of capital assumption.

There is, of course, an issue about "what Sraffa really intended." Here I shall merely rest my case with a quotation from Meek (1967, pp. 177–78):

> What both economists [Marx and Sraffa] are trying to show, in effect, is that (*when wages are given*) [italics added] the average rate of profits, and therefore the deviations of price ratios from embodied labour ratios, are governed by the ratio of direct to indirect labour in the industry whose conditions of production represent a sort of "average" of those prevailing over the economy as a whole. Marx reached this result by postulating as his "average" industry one whose "organic composition of capital" was equal to the "social average." But his result could only be a provisional and approximate one, since in reaching it he had abstracted from the effect which a change in the wage would have on the prices of the means of production employed in the "average" industry. Sraffa shows that the same result can be achieved, without abstracting from this effect at all, if we substitute his "standard" industry for Marx's industry of "average organic composition of capital." Sraffa's "standard" industry, seen from this point of view, is essentially an attempt to *define*

> "average conditions of production" in such a way as to achieve the identical result which Marx was seeking.

I shall now turn to why such arguments cannot be accepted.

IV. Five Fundamental Economic Objections to Sraffa's Standard Commodity

The economic usefulness of Sraffa's standard commodity depends on whether or not his new wage measure (w^s) is an appropriate concept in a variety of different situations. In this section I shall discuss five reasons why w^s is an inappropriate measure of any "real wage," except under highly unlikely circumstances:

1. Sraffa's weights used to construct his basket of goods are seen to be determined completely from the technology without regard for *consumer preferences.*

As before, $r^* \equiv (1/\lambda^*) - 1$, where λ^* is the Frobenius root of the matrix a and r^* is the maximal interest or profit rate at which the economy is productive enough to generate nonnegative steady-state prices. Associated with λ^* there is an eigenvector of the matrix a, which I denote by x^*. Sraffa's consumption basket weights $c^* = (c^*, \ldots, c^*)$ are defined by

$$c^* \equiv x^* - ax^* \tag{14}$$

(see, e.g., Burmeister 1968; Miyao 1977).

Meek's quoted phrase "when wages are given" must
to mean "when given wages are measured in terms
weights defined through (14) by Sraffa's standard
the wage rate measured w^s is not related in any wa
or preferences. Sraffa's unique consumption baske
termined from the technology alone; they are the
from the right-hand characteristic vector associat
benius root of the production technique matrix a. Ther
nomic reason why these weights should be relevant for defining any real wage.

Thus, for example, Sraffa's standard commodity may require that we assign relatively large weights in his consumption basket to commodities such as steel beams, which are never consumed by humans. To argue that "the deviations of price from embodied labour ratios" (Meek 1967, p. 177) are "determined" when wages, assumed given, are calculated using these special weights is an absurd proposition because, except in completely freak cases, any consumption basket relevant for an actual economy will bear no resemblance to Sraffa's weights that define his real wage w^s (also see Burmeister [1968, 1975*a*, 1980*a*], where I previously have called into question the economic significance of Sraffa's exercise). (We note, however, that in this model *all* real wage indices decrease with increasing r because W/P_i decreases as r increases, for all i.)

2. A unique standard commodity does not exist when there is more than one production technique available.

Let c^*_α denote the vector of Sraffa's consumption basket weights when production technique α is employed. Obviously, when many production techniques, say a, b, . . . , are technologically feasible, there exist as many corresponding sets of weights c^*_a, c^*_b, . . . , and "Sraffa wage" measures w^s_a, w^s_b, ...

Accordingly, unless only one production technique is optimal at all (feasible) interest or profit rates, the appropriate w^s will vary with r. This means that the linearity in figure 2 is destroyed; in its place we will find a series of line segments. It follows that "rates of exploitation" related to the slope of the w^s versus r trade-off will vary with r.

Likewise, joint production must be excluded. When intrinsic joint production exists, as in traditional wool-mutton examples, there may exist many steady-state equilibria price vectors at the same profit rate, depending on the pattern of final demand for commodities (see Burmeister and Kuga 1970). Since Sraffa ignores consumer preferences completely, in such joint production cases no unique w^s exists, but rather there exist an infinite number of analogous w^s measures depending on the pattern of final demand.

3. No primary factors other than labor can be admitted.

In essence Sraffa has provided a labor cost of production theory since only the production side of the economy is considered.

Once it is realized that it is the cost of production that is at issue, one must question the choice of labor as the unique factor of production used to measure "value." In general, the vector of labor input coefficients $(a_{01}, \ldots, a_{0n})$ must be replaced by the matrix

$$\begin{pmatrix} b_{11} & \cdots & b_{1n} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mn} \end{pmatrix},$$

where b_{ij} designates the quantity of the ith primary factor ($i = 1, \ldots, m$) required to produce one unit of commodity j ($j = 1, \ldots, n$). If $\mathbf{w} = (w_1, \ldots, w_m)$ is a vector of primary factor prices, in place of (4) we have

$$p(w, r) = wb[1 - (1 + r)a]^{-1}, \quad 0 \leq r < r^*. \qquad (15)$$

In particular, note that this criticism includes the obviously realistic case in which there are many types of heterogeneous labor inputs since these different types of labor inputs may be included in the b matrix and their corresponding wage rates in the vector **w**. Even at $r = 0$, at best one could have a valid "primary factor theory of cost," and that would require knowledge of relative primary factor prices. Since these primary factor prices are determined in general equilibrium and are not independent of demand, such a "primary factor theory of cost" must be rejected (see Burmeister 1975*b*). This observation is absolutely fatal for all attempts to construct a theory of "value" based on labor alone, or a theory of value that is independent of the demands and preferences of economic agents. That realistic economies are simultaneous general equilibrium systems is a complicating fact we cannot deny.

4. Even when there is only one technique of production and labor is the only primary factor, generally Sraffa's standard commodity does not exist when there are durable capital goods; only a subset of the possible depreciation patterns will admit a standard commodity.

The first part of Sraffa's 1960 analysis is based on a discrete-time circulating capital model where every capital good is completely used up during each production period. In University of Virginia lecture notes I have shown how Sraffa's results can be generalized to the case of exponential depreciation. Moreover, I allow depreciation rates to depend on use, so that the depreciation rate for steel can be larger in the machine tool industry than in the candy industry. Essentially, these generalizations are proved by considering net outputs and showing that the equilibrium pricing relationships can be predicted by another unique circulating capital model. The interested reader is referred to Samuelson (1983), where my results are discuss
tail.

Unfortunately, however, the exponential deprec
is only one of a subset of special cases for which
commodity exists. For many patterns of capital goo
some interest rates the price of a used machine will
the production techniques employed to produce
change with the interest rate. This fact means that objec
applies when durable capital goods exist, although for a differen
reason. These and other issues are discussed in a series of papers by Samuelson (1983, 1984), Baldone (1984), and Burmeister and Samuelson (1984); they are not central to the wage versus profit rate trade-offs derived below in Sections V and VI. The concluding paragraph of Samuelson (1983) summarizes the problem: "The complexities are in the nature of things, not a fabrication of the equilibrium analyst. Only the modern paradigm of general equilibrium—as begun by Smith and Ricardo and as improved by Walras, Fisher, Pareto, Dimitriev, von Bortkiewicz, von Neumann, Debreu, Arrow, Leontief,

and Sraffa—can do justice to the realities of technology and arbitrage."

5. Steady-state comparisons ignore the relevant intertemporal economic alternatives.

Both Sraffa's standard commodity and the Ricardo-Marx labor theory of value involve comparisons of alternative steady-state equilibria. Thus, for example, different points along the w-r line in figure 2 represent different steady states. These points do *not* accurately reflect the set of feasible alternatives open to an economy that happens to find itself in some particular steady-state equilibrium. "Comparative dynamics" does tell us how the original steady state compares with alternative steady states, but the set of possible steady states does not constitute the relevant choices open to an economy. Rather, starting from a particular steady state or from any other point, the set of technologically feasible dynamic paths (all having the same initial conditions) constitutes the feasible choices open to an economy.[5] Some of these feasible paths may converge to a steady state in either finite or infinite time, but in actuality such convergent paths constitute only a proper subset of the feasible options. In any event, an economy cannot move instantaneously from an initial steady-state equilibrium to another steady-state equilibrium, except perhaps in freak cases; a movement from one steady state to another involves a transition path of at least one time period duration. Thus labor theories of value restricted to steady states do not capture the relevant choices open to an economy.

V. Feasible Alternatives and Wage Rate Determination

We shall now free ourselves of the steady-state straitjacket and examine some economic properties of the relevant dynamic choices open to an economy. By so doing we can sketch a realistic mechanism of wage rate determination in competitive economies.

Because our interest is in what determines the real wage rate, we shall consider models with a single primary factor, labor; otherwise it would be necessary to consider a more complex model as outlined in remark 3 of the previous section. Our focus is on relevant intertemporal options open to an economy, and for expositional convenience the discussion initially will be limited to a one-sector economy. However, as we shall see in the next section, the results generalize to the case of heterogeneous commodities.

The technology is exceedingly simple. Inputs of k^t and l^t at the beginning of period t produce output y^{t+1} at the beginning of period

---

[5] This point has been recognized by Robinson (1975).

$t + 1$ (or end of period t). Output is divided into consumption and capital input during period $t + 1$:

$$y^{t+1} = c^{t+1} + k^{t+1}. \tag{16}$$

The production set consists of the set of feasible vectors

$$S^t = \{(y^{t+1}, k^t, l^t)\}. \tag{17}$$

The usual assumptions are made about the properties of S^t.[6]

Present value commodity and factor prices are denoted by p^t and w^t, respectively. Thus the condition for (discounted) profit maximization in period t is simply that the present value of outputs minus the present value of inputs be maximized at given (p^{t+1}, p^t, w^t), that is, max $(p^{t+1} \cdot y^{t+1} - p^t \cdot k^t - w^t \cdot l^t)$.

A feasible path $\bar{P}$ is termed competitive if there exist prices (p^{t+1}, p^t, w^t), for all t, such that $p^t \geqq 0$ and $p^t \neq 0$ for all t (nonnegative commodity prices); $w^t \geqq 0$ and $w^t \neq 0$ for all t (nonnegative primary factor prices); the primary factor is fully utilized ($l_t \equiv 1$); and along the path $\bar{P}$ the associated vectors $(\bar{y}^{t+1}, \bar{k}^t, \bar{l}^t)$ are profit maximizing at these prices; that is, for all t and for all vectors $(y^{t+1}, k^t, l^t) \in S^t$ we have

$$p^{t+1} \cdot \bar{c}^{t+1} + p^{t+1} \cdot \bar{k}^{t+1} - p^t \cdot \bar{k}^t - w^t \cdot \bar{l}^t \geqq p^{t+1} \cdot y^{t+1} - p^t \cdot k^t - w^t \cdot l^t. \tag{18}$$

Competitive paths as defined are nothing but dynamic paths along which the economy is always in general equilibrium, and steady states are but extremely special cases.[7]

In general, what economic properties are exhibited by
paths? I shall examine the following experiment.
omy following a competitive path P. In period t th
This capital input during period t yields output y^t
alternative feasible path $\bar{P}$ that is identical to P exc

---

[6] The following typical assumptions are made about the pr
tion set S^t, they are stated in such a way that y^{t+1}, k^t, and l^t m

ASSUMPTION 1 There exists some $(y^{t+1}, k^t, l^t) \in S^t$ with y^{t+1} 0 (pos
is possible), and if $(y^{t+1}, 0, 0) \in S^t$, then $y^{t+1} = 0$ (no positive output with ze

ASSUMPTION 2: S^t is a closed, convex set containing the origin (no increasing returns and $(0, 0, 0) \in S^t$

ASSUMPTION 3. If $(y^{t+1}, k^t, l^t) \in S^t$, $\bar{k}^t > k^t$, and $\bar{l}^t > l^t$, then there exists some $(\bar{y}^{t+1}, k^t, l^t) \in S^t$ such that $\bar{y}^{t+1} > y^{t+1}$ (capital and the primary factor are productive)

ASSUMPTION 4. If $(y^{t+1}, k^t, l^t) \in S^t$, then for any $\lambda > 0$ we also have $(\lambda y^{t+1}, \lambda k^t, \lambda l^t) \in S^t$ (constant returns to scale).

ASSUMPTION 5: If $(y^{t+1}, k^t, l^t) \in S^t$, $\bar{k}^t \geq k^t$, $\bar{l}^t > l^t$, and $y^{t+1} \geq \bar{y}^{t+1} > 0$, then $(y^{t+1}, \bar{k}^t, \bar{l})$ $\in S^t$ (free disposal).

[7] One should be careful, however, to avoid the false inference that such paths are consistent only with free competition. Competitive paths are consistent with the ordinary notion of free competition with perfect foresight, but other institutional forms (e.g., central planning) also are consistent with the purely technical definition of a competitive path.

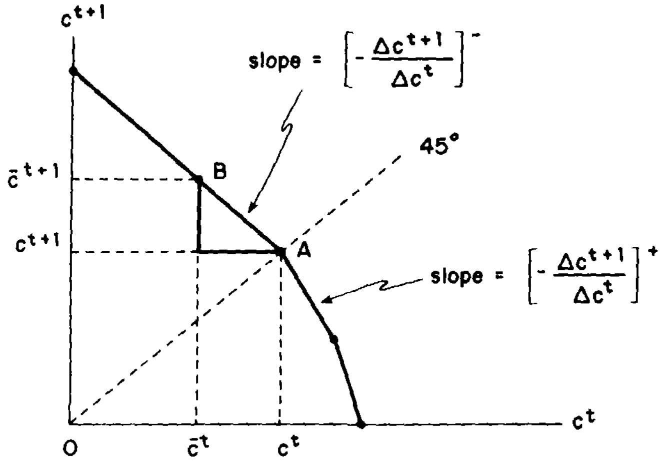

c^{t+1}
slope = $\left[-\frac{\Delta c^{t+1}}{\Delta c^t}\right]^-$
45°
$\bar{c}^{t+1}$
B
c^{t+1}
A
slope = $\left[-\frac{\Delta c^{t+1}}{\Delta c^t}\right]^+$
c^t
O
$\bar{c}^t$
c^t

FIG. 3

$\neq \bar{c}^t$, $k^t \neq \bar{k}^t$, and $c^{t+1} \neq \bar{c}^{t+1}$ That is, I am assuming that consumption along the two paths is identical for all time periods except in periods t and $t + 1$, while the capital stocks are identical for all time periods except period t. (The competitive path P may represent a steady state, but this case is very special.) Economic common sense suggests that if $\bar{c}^t$ is lower than c^t, then this "consumption sacrifice" will enable the economy to enjoy a $\bar{c}^{t+1}$ larger than c^{t+1}. Indeed, not only is this intuition correct, but also the ratio of the consumption differences in periods $t + 1$ and t provides bounds on the ratio of present value prices.

To make matters precise, we must define the rate of intertemporal (consumption) transformation between periods t and $t + 1$. The left-hand marginal rate of intertemporal transformation between periods t and $t + 1$ is defined as

$$\lim_{\Delta c^t \to 0^+} \left(-\frac{\Delta c^{t+1}}{\Delta c^t}\right) \equiv \left(-\frac{\Delta c^{t+1}}{\Delta c^t}\right)^+, \quad \Delta c^t \equiv c^t - \bar{c}^t. \tag{19}$$

Note that this limit is defined as Δc^t approaches zero through positive values, and hence it is the "consumption sacrifice" case in which $\bar{c}^t < c^t$. Similarly, the right-hand marginal rate of intertemporal transformation between periods t and $t + 1$ is defined as

$$\lim_{\Delta c^t \to 0} \left(-\frac{\Delta c^{t+1}}{\Delta c^t}\right) \equiv \left(-\frac{\Delta c^{t+1}}{\Delta c^t}\right)^-. \tag{20}$$

Right- and left-hand marginal rates of transformation are illustrated in figure 3. The economy is initially at point A on path P with consumption c^t and c^{t+1}. However, the alternative feasible path $\bar{P}$

allows a "consumption sacrifice" with $\bar{c}^t < c^t$, and the economy then benefits by having $\bar{c}^{t+1} > c^t$. Thus the point B in figure 3 is on the alternative feasible path $\bar{P}$, and the slope of the line segment AB is the left-hand marginal rate of intertemporal transformation.

The feasible alternatives for this experiment can be described by a production function

$$c^{t+1} + k^{t+1} = G(k^t + c^t, k^t, k^{t-1}, l^t, l^{t-1}), \tag{21}$$

which is concave and homogeneous of degree one.[8] Along any feasible path $\bar{P}$, $dk^t + dc^t = 0$ and $dk^{t+1} = dk^{t-1} = dl^t = dl^{t-1} = 0$. Thus when $G(\cdot)$ is differentiable,

$$dc^{t+1} + 0 = G_1 \cdot 0 + G_2 dk^t + 0 + 0 + 0$$

or

$$-\frac{dc^{t+1}}{dc^t} = G_2 = \frac{p^t}{p^{t+1}} \equiv 1 + r^t,$$

where r^t is the (own) interest rate over period t. When $G(\cdot)$ is not differentiable, it follows from the well-known relationship between directional derivatives and the subgradient set that

$$(1 + r^t)_{\min} \equiv \left(\frac{\Delta c^{t+1}}{\Delta c^t}\right) \leq \frac{p^t}{p^{t+1}} = 1 + r^t \leq \left(\frac{c^{t+1}}{c^t}\right)^{+} \equiv (1 + r^t)_{\max}. \tag{22}$$

The latter provides the fundamental bounds on competitive [illegible] and hence the interest rate for the path $\bar{P}$.[9]

Now let the real wage rate w^t/p^{t+1} be denoted [illegible] differentiable, the real wage rate is given by

$$\frac{\partial y^{t+1}}{\partial l^t} = \frac{\partial(c^{t+1} + k^{t+1})}{\partial l^t} = G_3 = \frac{w^t}{p^t}.$$

When $G(\cdot)$ is not differentiable, it again follows from [illegible] concave functions that the real wage rate is bounded

$$\omega^t_{\min} \leqq \omega^t \leqq \omega^t_{\max}.$$

Since $G(\cdot)$ is homogeneous of degree one, in all cases there are [illegible]

---

[8] I am grateful to a referee for suggesting the simple proofs below which are immediate consequences of the properties of this production function.

[9] In other words, in the space of feasible consumption at times t and $t + 1$ the slope of the trade-off need not be the same from each direction, as at point A in fig. 3; the competitive rate of interest is thus bounded by these slopes.

[10] I assume that w^t is a post factum wage rate used to purchase goods produced during period t at prices p^{t+1}; hence w^t/p^{t+1} is the correct measure of the real wage rate. Were wages paid at the beginning of period t for the purchase of goods produced during period $t - 1$ at prices p^t, nothing essential would be altered except that the straight line segment in fig. 4 would be replaced by a negatively sloped curve.

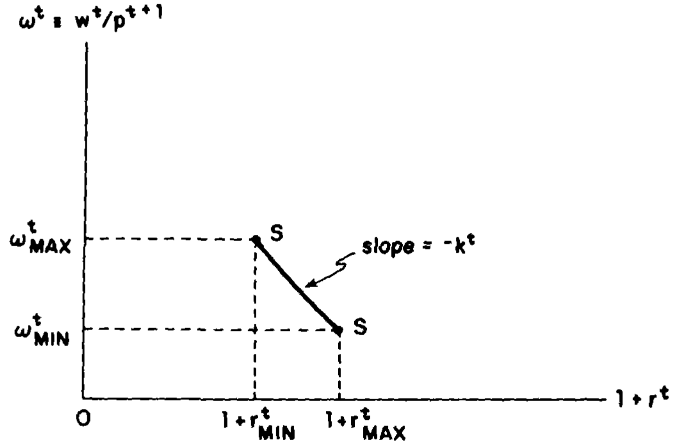


FIG. 4.—When the possibility of real wage rate indeterminacy exists, the marginal rate of intertemporal transformation provides technologically determined bounds on the interest rate that bound the values of the real wage rate that are consistent with competitive equilibrium, independently of whether or not the economy is in a steady-state equilibrium

profits along every competitive path. Thus even when the real wage rate is determined only up to the bounds given by (24), it is still true that $\omega^t + (1 + r^t)k^t$ = constant and

$$\frac{d\omega^t}{dr^t} = -k^t. \tag{25}$$

The segment of the line labeled SS in figure 4 represents this tradeoff between the interest rate and the real wage rate when a point such as A in figure 3 is selected. It holds even when the economy is not in a steady-state equilibrium.[11]

VI. The Case of Heterogeneous Capital Goods

The existence of heterogeneous capital goods is known to destroy many simple neoclassical parables, and it is important to establish that our essential conclusions do not depend on a one-sector formulation. The production function for our experiment can be described by

$$c^{t+1} = F(c^t, k^{t-1}, l^t, l^{t-1}), \tag{26}$$

where k^t is now interpreted as a vector of heterogeneous capital goods $(k^t_1, \ldots, k^t_n)$.[12] Similarly, now $p^t = (p^t_1, \ldots, p^t_n)$, while the price of the single consumption good c^t is denoted by p^t_c.

---

[11] In a steady-state equilibrium with $c^{t+1} = c^t \equiv c$, $k^{t+1} = k^t \equiv k$, $\omega^t = \omega^{t+1} \equiv \omega$, and $r^{t+1} = r^t \equiv r$, we have $\omega = w^t/p^{t+1} = c - rk$, $d\omega/dr = -k$ Also see n. 12 below.

[12] Note that $F(\cdot)$ has been defined with k^t netted out.

The own rate of return for the ith capital good is defined by

$$r_i^t \equiv \frac{p_i^t}{p_i^{t+1}} - 1.$$

Except in a steady-state equilibrium, the n own rates of return and the own rate of return for the consumption good, $r^t \equiv (p_c^t/p_c^{t+1}) - 1$, need not be equal. However, by an argument exactly analogous to that in the previous section,

$$\frac{\partial \omega^t}{\partial r_i^t} = -\frac{p_i^{t+1}}{p_c^{t+1}} k_i^t, \quad i = 1, \ldots, n, \tag{27}$$

where the partial derivative is taken with the price ratios p_i^{t+1}/p_c^{t+1} held constant for all $i = 1, \ldots, n$.[13] Thus in general we see that there is a negatively sloped linear relationship between the real wage rate and the own rate of return on every capital good.

The assumption of steady-state equilibrium simplifies matters in this heterogeneous capital case, for then $c^t = c^{t+1} \equiv c$, $\omega^t = \omega^{t+1} \equiv \omega$, $k_i^t = k_i^{t+1} \equiv k_i$, and $r_i^t = r^t = (p_c^t/p_c^{t+1}) - 1 \equiv r$ for all $i = 1, \ldots, n$; that is, in an efficient steady-state equilibrium quantities are constant and all $n + 1$ own rates of return are equal. Thus, using $p_i^t = p_i^{t+1}(1 + r^t)$ and

$$\sum_{i=1}^{n} \frac{p_i^{t+1}}{p_c^{t+1}} k_i^t = \sum_{i=1}^{n} \frac{p_i^{t+1}}{p_c^{t+1}} k_i \equiv v,$$

it follows from the zero-profit condition that

$$\omega = c - rv.$$

Hence

$$\frac{\partial \omega}{\partial r} = -v,$$

where the partial derivative is taken keeping the [illegible] constant. Figure 4 immediately may be reinterpreted as the steady-state trade-off between the real wage rate and the real (consumption good) interest rate; the slope of this trade-off is $-v$ between $r_{\min}$ and $r_{\max}$.[14]

---

[13] We are still considering a choice of c^t at a "kink" such as point A in fig. 3; thus c^{t+1}, c^t, k_i^t, and k_i^{t+1} are all constant by assumption.

[14] Note that for any steady-state equilibrium with $c^t = c^{t+1} = c$, a real wage rate vs. profit rate trade-off exists if and only if there is a kink along the 45° line in fig. 3; otherwise r is uniquely determined by the slope of the frontier at the $c^t = c^{t+1} = c$ point. Accordingly, we see that such trade-offs may be observed more often out of steady-state equilibrium.

VII. Concluding Remarks

We have come a long way from Sraffa's analysis. In particular, (1) There is no restriction to steady-state equilibria; any competitive intertemporal path is allowed. (2) Technological change is admitted since the production set S^t can change over time. (3) Consumer preferences about intertemporal consumption have a role in determining the real wage rate or the real wage rate versus interest rate trade-off. (4) Joint production and heterogeneous capital goods are allowed.

Given this framework, I have identified situations in which legitimate interest rate versus real wage rate trade-offs exist. As illustrated by figure 4, classical and neoclassical economic intuition is confirmed: When a trade-off exists, lower real wage rates are associated with higher interest rates, and vice versa.

Most important, note that along a trade-off frontier such as SS in figure 4, no quantities change. This fact means that we have identified a genuine bargaining or conflict situation between workers, on the one hand, and owners of the capitalistic means of production, on the other. *Within this framework we can tell realistic tales about owners* of the means of production "exploiting" workers by offering the minimum real wage rate $\omega^t_{\min}$.

But there is a symmetry: workers may exercise "union power" by refusing to supply labor at real wage rates lower than $\omega^t_{\max}$. Clearly some type of game theory or "class struggle" solution is required to resolve this remaining indeterminacy.[15] Moreover, this indeterminacy is consistent with competitive paths and much of modern neoclassical economics. How important the issue is depends on the technology. As the technology becomes "smoother," with more and more feasible production techniques, the range of indeterminacy shrinks, and it vanishes in the limit if the technology is differentiable.

However, even if there is a short-run trade-off between wages and profits, it is not clear whether or not a higher wage rate at the expense of a lower profit rate can be sustained in the longer run. Although a more complete model would be required to examine this question in detail, presumably there would be a shift of capital resources away from sectors of the economy with lower profit rates to those with higher profit rates. Thus in the long run supply conditions might provide stricter bounds on the wage and profit rates consistent with competition than the short-run bounds derived in this paper.[16]

---

[15] The results of this paper may be interpreted as providing one reconciliation between perfect competition and some of Akerlof's recent theories about labor contracting (see, e.g., Akerlof 1982)

[16] I am grateful to Paul A. Samuelson for pointing out to me that long-run competition might shrink the wage vs. profit rate bounds derived above

I conclude, therefore, that workers and capitalists may stand in conflict, but (revolutions aside) competition restricts the outcomes to well-defined bounds, and these bounds may be very small, especially in the long run.

References

Akerlof, George A. "Labor Contracts as Partial Gift Exchange." *Q J E.* 97 (November 1982): 543–69.

Baldone, S. "Durable Capital Inputs: Conditions for Prices Ratios to Be Invariant to Profit-Rate Changes: A Comment to Professor Samuelson." *Zeitschrift für Nationalokonomie,* forthcoming (1984).

Brown, Murray, and Chang, Winston W. "Capital Aggregation in a General Equilibrium Model of Production." *Econometrica* 44 (November 1976): 1179–1200.

Bruno, Michael; Burmeister, Edwin; and Sheshinski, Eytan "The Nature and Implications of the Reswitching of Techniques." *Q J E* 80 (November 1966): 526–53.

Burmeister, Edwin. "On a Theorem of Sraffa" *Economica* 35 (February 1968): 83–87.

———. "A Comment on 'This Age of Leontief . . . and Who?'" *J Econ Literature* 13 (June 1975): 454–57. (*a*)

———. "Many Primary Factors in Non-Joint Production Economies." *Econ Record* 51 (December 1975): 486–512 (*b*)

———. "Real Wicksell Effects and Regular Economies" In *Essays in Modern Capital Theory,* edited by Murray Brown, Kazuo Sato, and Paul Zarembka Amsterdam: North-Holland, 1976.

———. *Capital Theory and Dynamics.* New York. Cambridge Univ Press, 1980 (*a*)

——— "Comments on Capital Aggregation." In *The Measurement of C*, edited by Dan Usher. Chicago: Univ Chicago Press (for N (*b*)

———. "Critical Observations on the Labor Theory of Val *Economics and Development: Essays in Memory of Ta Chung* rence R. Klein, Marc Nerlove, and Sho Chieh Tsiang Ne Press, 1980. (*c*)

Burmeister, Edwin, and Kuga, Kiyoshi "The Factor-Pric classical Multi-Sector Model." *Internat Econ. Rev* 11 (Feb 74.

Burmeister, Edwin, and Samuelson, Paul A. "Aspects of Joint Unpublished manuscript, Univ. Virginia, 1984.

Dorfman, Robert; Samuelson, Paul A ; and Solow, Robert M *Linear Programming and Economic Analysis.* New York. McGraw-Hill, 1958.

Eatwell, John L. "Mr. Sraffa's Standard Commodity and the Rate of Exploitation." *Q.J E* 89 (November 1975): 543–55.

Flaschel, Peter. "Actual Labor Values in a General Model of Production." *Econometrica* 51 (March 1983): 435–54

Harcourt, Geoffrey C. "The Sraffian Contribution: An Evaluation" In *Classical and Marxian Political Economy. Essays in Honour of Ronald L Meek,* edited by Ian Bradley and Michael Howard. London: Macmillan, 1982

Meek, Ronald L *Economics and Ideology and Other Essays Studies in the Development of Economic Thought.* London: Chapman & Hall, 1967

Miyao, Takahiro. "A Generalization of Sraffa's Standard Commodity and Its Complete Characterization." *Internat. Econ. Rev.* 18 (February 1977): 151–62.

Morishima, Michio. *Marx's Economics: A Dual Theory of Value and Growth.* London and New York: Cambridge Univ. Press, 1973.

Nell, Edward J. "Value and Capital in Marxian Economics." *Public Interest* (Special Issue, 1980), pp. 174–200.

Parys, Wilfried. "A Simple Criterion for Equal Value Composition of Capital." *Q.J.E.* 91 (August 1977): 511–12.

———. "The Deviation of Prices from Labor Values." *A.E.R.* 72 (December 1982): 1208–12.

Ricardo, David. *The Principles of Political Economy and Taxation.* London: Dent, 1911. Reprint ed. Homewood, Ill.: Irwin, 1963.

Ripley, Frank C., and Segal, Lydia. "Price Determination in 395 Manufacturing Industries." *Rev. Econ. and Statis.* 55 (August 1973): 263–71.

Robinson, Joan. "The Unimportance of Reswitching." *Q.J.E.* 89 (February 1975): 32–39.

Samuelson, Paul A. "Understanding the Marxian Notion of Exploitation: A Summary of the So-Called Transformation Problem between Marxian Values and Competitive Prices." In *The Collected Scientific Papers of Paul A. Samuelson*, edited by Robert C. Merton. Vol. 3. Cambridge, Mass.: M.I.T. Press, 1972.

———. "Durable Capital Inputs: Conditions for Price Ratios to Be Invariant to Profit-Rate Changes." *Zeitschrift für Nationalökonomie* 43, no. 1 (1983): 1–20.

———. "When Standard Commodities Do and Don't Exist." *Zeitschrift für Nationalökonomie*, forthcoming (1984).

Samuelson, Paul A., and von Weizsäcker, Carl C. "A New Labor Theory of Value for Rational Planning through Use of the Bourgeois Profit Rate." In *The Collected Scientific Papers of Paul A. Samuelson*, edited by Robert C. Merton. Vol. 3. Cambridge, Mass.: M.I.T. Press, 1972.

Sellekaerts, Willy, and Lesage, Richard. "A Reformulation and Empirical Verification of the Administered Prices Inflation Hypothesis: The Canadian Case." *Southern Econ. J.* 39 (January 1973): 345–60.

Sraffa, Piero. *Production of Commodities by Means of Commodities. Prelude to a Critique of Economic Theory.* London and New York: Cambridge Univ. Press, 1960.

Stigler, George J. "Ricardo and the 93% Labor Theory of Value." *A.E.R.* 48 (June 1958): 357–67.

Weiss, Leonard. "Quantitative Studies of Industrial Organization." In *Frontiers of Quantitative Economics*, edited by Michael D. Intriligator. Amsterdam: North-Holland, 1971.

Bequests and the Size of Population When Population Is Endogenous

Marc Nerlove

University of Pennsylvania

Assaf Razin and Efraim Sadka

Tel-Aviv University

The consequences of interfamily bequests and endogenous population size for optimal and competitive population sizes and bequests are explored when individuals' utility is a function of own consumption, the number of children, and the welfare of their children. When a bequest benefits each member of a second-generation family, we show that competition leads to underprovision for future generations. Under a separability assumption, the number of children is also shown to be too large.

I. Introduction

In this note we consider the problem of a utility-m
when utility depends on own consumption, number
indirectly, on children's welfare through the bequests
A novel feature of our approach is that the welfare
family depends on the bequest of both the child
spouse. Because each family derives utility from the beque

We wish to thank Elhanan Helpman, Murray Kemp, and John Mellor with whom we have had useful conversations on the topic of this note. This research was supported by grants from the National Institute on Aging to Northwestern University and the University of Pennsylvania and from the U.S.-Israel Binational Science Foundation to Tel Aviv University. Part of Nerlove's contribution was undertaken at the International Food Policy Research Institute.

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0006$01.50

families through marriages, Pareto optimality can be attained only if the parents are free to bargain with one another about what each child's family will leave to its children. This was true in biblical times when parents negotiated *Neduniahs* ("doweries") and *Mohars* ("bride prices") with one another. But in those days parents had certain property rights in their children. In modern times, when so much of parents' bequests is in the form of human capital, property rights such as these are difficult or impossible to enforce and marriages are not "arranged" in this manner by parents. If the children are free to choose and marry for "love," that is, on the basis of considerations unrelated to bequests, and if bequests are determined prior to the choice, then we show that the level of bequests will be less than a symmetric Pareto-optimal allocation. We also show that under a suitable separability assumption, the number of children each family bears is larger than a symmetric Pareto-optimal allocation.

We emphasize that here we consider only the question of allocative efficiency among members of the first generation (parents); that is, the social welfare is the individual parent's and the children enter only through their parents' utility functions. The implications of alternative social welfare functions have been considered elsewhere (Nerlove, Razin, and Sadka 1982*c*).

II. Formulation

Let there be two families in the current generation and only two generations. The problem may be formulated as follows: c_i = the consumption of the ith family, n_i = the number of children of the ith family, b_i = the per child bequest of the ith family, R_i = the resources available to the ith family for consumption and bequest, $i = 1, 2$.

The total bequest of two children who marry one another will be the sum of the bequests to each child, that is, $b_1 + b_2$. If each family's utility function, u, is identical and if each family is subject to the same resource constraint, R, each will behave in an identical manner, so the number of children available to marry each other will be identical (no singles and no incest). Thus, the ith family chooses c_i, n_i, and b_i so as to maximize

$$u(c_i, n_i, b_1 + b_2) \tag{1}$$

subject to the resource constraint

$$R \geq c_i + b_i n_i, \tag{2}$$

where b_j, $j \neq i$, is taken as a parameter by the ith family (parametric externalities). It is assumed that u is strictly increasing in consumption and bequests, but not necessarily in number of children.[1] A competitive (Nash) solution is a 6-tuple $(\bar{c}_1, \bar{n}_1, \bar{b}_1, \bar{c}_2, \bar{n}_2, \bar{b}_2)$ such that $(\bar{c}_i, \bar{n}_i, \bar{b}_i)$ solves the problem defined above, $i = 1, 2$. Because of the assumed symmetry, we have $\bar{c}_1 = \bar{c}_2 \equiv \bar{c}$, $\bar{b}_1 = \bar{b}_2 \equiv \bar{b}$, and $\bar{n}_1 = \bar{n}_2 \equiv \bar{n}$.

III. Pareto Optimality and Individual Maximization

A symmetric Pareto optimum $(c^*, n^*, b^*, c^*, n^*, b^*)$ is obtained by a simultaneous choice of $(c_1, n_1, b_1, c_2, n_2, b_2)$ so as to maximize

$$u(c_1, n_1, b_1 + b_2) + u(c_2, n_2, b_1 + b_2) \tag{3}$$

subject to

$$2R \geqslant c_1 + b_1 n_1 + c_2 + b_2 n_2 \tag{4}$$

and subject to the symmetry condition $c_1 = c_2$, $n_1 = n_2$, $b_1 = b_2$. This amounts to finding a Pareto-optimal solution in which each family is treated equally.[2]

A straightforward comparison between the necessary first-order conditions of (1)–(2) and those of (3)–(4) reveals that the two solutions, the competitive and the symmetric optimal solutions, differ (provided we are not at a corner). Observe that if, in the competitive case, marriages are "arranged" in such a way as to be a symmetric solution to a cooperative bargaining game (non-Nash), there will be no difference between the competitive and socially optimal solutions

IV. Bequests in the Two Cases

Since in the competitive solution each family ignores its
to the welfare of the other family through its beq
marrying the children of the other family, one wo
little is bequeathed in the competitive solution as

---

[1] The utility function need not be increasing everywhere in n; it is
struct plausible utility functions that are not. For example, $u(c_i, n_i, b_1 + b_2) = u[c_i/(1 + n_i), n_i, b_1 + b_2]$; i.e., utility depends on per capita family consumption in the first generation (see Nerlove, Razin, and Sadka 1982*b*). At the internal solution, u will always be increasing with n since children have a positive cost $b > 0$; see eq. (2).

[2] This symmetric Pareto optimum is also a symmetric social optimum when the children's welfare enters the social objective function only as an argument in the parents' utilities and not independently; see Nerlove et al. (1982*a*) for alternative social welfare functions. A solution to maximizing (3) subject to (4) will not necessarily be symmetric, without imposing further conditions, because (4) is not convex in b and n and u is not necessarily concave.

symmetric optimum, that is, $\bar{b} \leq b^*$. This is indeed the case.[3] To prove this, suppose to the contrary that $\bar{b} > b^*$. Then

$$u(c^*, n^*, b^* + b^*) - u(c^*, n^*, b^* + \bar{b}) < 0, \tag{5}$$

because u is strictly increasing in b. But

$$u(c^*, n^*, b^* + \bar{b}) - u(\bar{c}, \bar{n}, \bar{b} + \bar{b}) < 0, \tag{6}$$

because (c^*, n^*, b^*) satisfies (4) and hence also (2), and $(\bar{c}, \bar{n}, \bar{b})$ maximizes u subject to (2). Adding (5) and (6) yields

$$u(c^*, n^*, b^* + b^*) < u(\bar{c}, \bar{n}, \bar{b} + \bar{b})$$

in contradiction to the Pareto optimality of (c^*, n^*, b^*).

V. Number of Children in the Two Cases

In this section we compare the number of children in the two solutions. Notice that the bequest b is the "price" of children in the budget constraint (2) or (4) (see Becker and Lewis 1973). The implications are as follows: First, since b is lower in the competitive solution, children are essentially "cheaper." Second, there is a lower real income (welfare) in the competitive solution, due to the externality; thus, if children are a normal good, the income effect tends to counteract the price effect. Third, since, in general, the level of bequests affects the marginal rate of substitution between children and family consumption, the projection of the indifference map in the c-n plane must shift. Therefore, it is impossible to draw any general conclusions about the relationship between $\bar{n}$ and n^*.[4]

When there is weak separability between (c, n) and b, there is no shift in the indifference map for consumption and children. In this special case the price effect dominates and we can show that the competitive number of children is larger than the optimal one ($\bar{n} > n^*$). To see this observe that weak separability means that $u(c_t, n_t, b_1 + b_2)$ may be written as $U[\phi(c_t, n_t), b_1 + b_2]$. Therefore, (c^*, n^*) must maximize $\phi(c, n)$ subject to $R \geqslant c + nb^*$, while $(\bar{c}, \bar{n})$ must maximize $\phi(c, n)$ subject to $R \geqslant c + n\bar{b}$. Assuming that ϕ is not such that n is a

---

[3] The reader who is familiar with the literature on externalities would notice that this result follows from the assumed symmetry in this note (see Buchanan and Kafoglis 1963, Diamond and Mirrlees 1973; Sadka 1978).

[4] Distributional criteria related to bequests are discussed in Nerlove et al (1982c) There we show that, in the absence of an externality associated with bequests, maximization of a Benthamite social welfare function (sum of parents' and children's utilities) will lead to an optimal population always larger than a Millian social welfare function (per capita utility of parents and children). However, the laissez-faire solution (maximization of parents' utilities only) may lead to a population larger than the Benthamite optimum or smaller than the Millian optimum.

Giffen good, it follows that $\bar{n} > n^*$ because $\bar{b} < b^*$. Without the assumption of weak separability, it is easy to find examples of plausible utility functions for which the socially optimal number of children is less than the individually optimal number. For example, if utility is of the form $U[c_t, \min(b_1 + b_2, n_t)]$, $\bar{n} < n^*$ because $\bar{b} < b^*$, and there are fixed proportions between bequests and the number of children.

VI. Concluding Remarks

In this note we identify a source of market failure in the provision of parents for their children. If parents do not take into account the effect of their bequests on the welfare of families to whom they are potentially related by the marriage of their children, they will bequeath too little. In this case bequests should be subsidized on efficiency grounds rather than taxed. Under the assumption of weak separability of bequests in the parents' utility function from family consumption and number of children, the less than optimal level of bequests is associated with too large a rate of population growth.

We are concerned here only with market failure due to the potential externality introduced by bequests. The optimal population size and level of bequests under alternative social welfare functions have been considered elsewhere (Nerlove et al. 1982*c*).

Finally, we note that the externality identified here applies not only to bequests in the form of physical capital but, especially, to bequests in the form of investments in the human capital of children. It is the latter that effectively prevent establishment of a system of property rights of parents in children that would eliminate the externality

References

Becker, Gary S., and Lewis, H. Gregg. "On the Int[illegible] Quantity and Quality of Children." *J.P.E.* 82, no. [illegible] 1973): S279–S288

Buchanan, James M., and Kafoglis, Milton Z. "A Note [illegible] ply." *A.E.R.* 53 (June 1963): 403–14

Diamond, Peter A., and Mirrlees, James A. "Aggregate Pr[illegible] sumption Externalities." *Q.J.E.* 87 (February 1973): 1–2[illegible]

Nerlove, Marc; Razin, Assaf, and Sadka, Efraim. "Population Size and [illegible] Social Welfare Functions of Bentham and Mill." *Econ. Letters* 10, nos. 1, 2 (1982): 61–64. (*a*)

———. "Child Allowances, Optimal Population Policy, and Intergenerational Justice." Working Paper no. 9-82. Tel Aviv: Tel-Aviv Univ., Foerder Inst. Econ. Res., 1982. (*b*)

———. "Population Size: Individual Choice and Social Optima." Working Paper no. 22-82. Tel Aviv: Tel-Aviv Univ., Foerder Inst. Econ. Res., 1982. (*c*)

Sadka, Efraim. "A Note on Aggregate Production with Consumption Externalities." *J. Public Econ.* 9 (February 1978): 101–5.

Comments

The Decline in Male Labor Force Participation: Comment

Robert H. Haveman and Barbara L. Wolfe

University of Wisconsin—Madison

In an article in this *Journal*, Parsons (1980*a*) analyzes the doubling in the rate of nonparticipation in the labor force of older men since 1946. Using cross-sectional data from the National Longitudinal Survey (Parnes et al. 1970) in an index function model of the labor force participation decision, he tests the assertion that the withdrawal of older males is attributable to the rapid expansion of alternatives to work provided by the Social Security Disability Insurance (SSDI) program. He reports an elasticity of nonparticipation with respect to the replacement rate of SSDI benefits of .63[1] and concludes that the recent decline in labor force participation can be "largely explained by the increased generosity of social welfare transfers, particularly Social Security disability payments" (p. 130). The magnitude of Parsons's estimate and its impact on the public debate over disability transfer policy require careful scrutiny of its basis. In Section I, we describe Parsons's econometric model and construction of variables; in Section II we provide a critique. In Section III, we test the robustness of Parsons's elasticity estimate by presenting alternative estimates based on corrections and extensions of his basic model. Section IV concludes.

The able research assistance of Mitchell Post and the econometric help of Arthur Goldberger are gratefully acknowledged. David Betson, Sheldon Danziger, Chris Flinn, Arthur Goldberger, Jacques van der Gaag, James Heckman, Jennifer Warlick, Aaron Krute, Jesse Levy, and an anonymous referee of this *Journal* provided helpful comments.

[1] In a companion paper (Parsons 1980*b*), Parsons estimates an elasticity triple this value using the same data base but a different time period and a slightly different econometric specification.

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0007$01.50

I

Parsons employs a probit version of a discrete choice econometric model of the form

$$LFP_i = f(SSB_i/W_i, MORT_i, WEL_i/W_i, UR_i, AGE_i, \epsilon_i),$$

where LFP_i is a dummy variable equal to one if individual i is a labor force participant and equal to zero if individual i is not a labor force participant; SSB_i is social security benefits available to individual i; W_i is i's hourly wage rate; WEL_i is the annual welfare payment in i's state divided by the number of families in i's state with annual income less than $3,000; $MORT_i$ is an index of the mortality experience of i, taken to represent health status; UR_i is the fraction of the year i spent unemployed; AGE_i is i's actual age in the year considered; and ϵ_i is a normally distributed error term on observation i. Both a linear form of this model and a model allowing for interactions between $MORT_i$ and SSB_i/W_i and WEL_i/W_i are fitted.

The basic data set contains observations on 3,219 males aged 45–59 in 1966; the dependent variable is the labor force status of these men in 1969. The 1966 values of the primary variables are used to reduce censoring problems, although those observations without a reported 1966 wage rate are apparently excluded. W_i represents the 1966 value scaled up to the 1969 level by a factor representing the growth in per capita income over the 3-year period. Parsons estimated SSB_i from extrapolated values of W_i using the 1969 statutory formula for primary SSDI benefits. The variable $MORT_i$, taken to represent disability status, is a weighted average of mortality dummy variables based on subsequent mortality in each of the years 1969–76. UR is a 1966 value, AGE_i a 1969 value.

Column 1 of table 1 reports Parsons's results. Al
are significant and of the expected sign. The coeff
taken by Parsons to confirm his hypothesis regard
effects of SSDI benefits and serves as the basis f
mate, which is also reported in the table

II

For several reasons, we believe that Parsons's results are unreliable and that his large reported elasticity is misleading.[2] First, Parsons's replacement rate variable (SSB/W) confounds the separate impacts that labor market and disability transfer options have on the partici-

---

[2] The −.03 elasticity estimate in table 1 is the response in terms of participation with respect to the replacement rate; it is equivalent to the .63 response for nonparticipation reported by Parsons

TABLE 1

Probit Estimates of the Labor Force Participation of Men Aged 48–62
(Data for 1969 in Parsons, for 1977 in H-W Estimates)

	Parsons (1)	H-W Replication (2)	Column 2 Specification with Split Replacement Rate (3)	Column 2 plus Inclusion of Dependent Benefits and Additional Controls[3] (4)	Column 2 plus Selectivity Correction (5)	Column 3 with Eligibility Adjustment (6)	Column 3 plus Selectivity Correction Eligibility Adjustment, and Additional Controls (7)
Constant	5.626 (10.94)*	6.094 (4.97)	5.954 (4.42)	6.137 (4.79)	5.673 (5.53)	5.889 (4.68)	5.83 (5.20)
Price variables:							
SSB/W	−.00542 (−2.48)	−.00626 (−1.59)	.	. .	−.0032 (−.85)	. . .	. .
SSBD/W	. .	.	. .	−.0017 (−.43)	. .	. . .	. . .
SSB	. . .	. . .	−.0007 (−.41)		. .	. . .	. .
PSSB	. .	.	. .		.	−.0024 (−.87)	−.0012 (−.48)

W	...		.022		.	.022	−.019
			(.57)			(.78)	(−.67)
WEL/W	−.00278	.00260	.	.020	.0014	...	...
	(−1.97)	(1.72)		(1.28)	(1.13)		
WEL	...	...	.0003	...	...	.0004	−.0001
			(.87)			(1.00)	(−.31)
UR	−.835	−.457	−.498	−.364	−.178	−.524	−.1858
	(−3.13)	(−.73)	(−.79)	(−.31)	(−.31)	(−.83)	(−.01)
Demographic variables:							
Age	−.0639	−.0683	−.0700	−.0700	−.0643	−.073	−.0701
	(−7.15)	(−3.09)	(−3.18)	(−3.03)	(−3.47)	(−3.25)	(−3.51)
Disability status†	−.944	−2.008	−2.017	−2.080	−2.331	−1.621	−2.54
	(−8.21)	(−9.88)	(−9.89)	(−13.78)	(−13.46)	(−3.32)	(−4.99)
Elasticity of labor force participation with respect to social security benefits	−.030‡	−.0205	−.0147	−.0069	−.0177	−.0067	−.0056
−2 × log likelihood ratio	1,301	143.6	141.0	141.5	264.0	141.6	272.0
N	3,219	741	741	741	796	741	796

---

*Asymptotic normal test statistic reported in parentheses.

†The constructed subsequent mortality index for Parsons' index of extent of reported disability for H-W.

‡Parsons reports an elasticity of nonparticipation of .63, which is equal to the participation elasticity of −.030.

§The additional control variables added are (1) a dummy variable where not married, no children = 1, (2) a dummy variable where married, no children = 1, (3) other household income, and (4) a dummy where long-term occupation is manager or professional = 1. The coefficients and t-values for these variables in the col. 3 specification are, respectively, −3.46 (−1.07), −.96 (−0.46), −.032 (−1.76), .151 (0.05). These coefficients and asymptotic t-values remain stable throughout the remaining runs reported.

pation decision. While true expected disability benefits depend on a variety of factors such as severity, number of dependents, other available programs, and work history, the numerator of Parsons's replacement rate (expected primary beneficiary SSDI payments)[3] is a monotonic function of, and hence highly collinear with, the denominator (W). By neglecting other factors determining expected disability transfers, Parsons has specified a model in which labor force participation depends on a proxy for the wage rate divided by the wage rate itself. Has he found that high expected disability benefits reduce participation or merely that high expected wages increase it?

Our second criticism concerns Parsons's choice of the disability transfer income variable (SSB), which is the numerator of the replacement rate. Ideally, in modeling the work/nonwork options confronting individuals in a utility (income) maximization framework, it is desirable to specify the full, nonlinear structure of the budget constraint they confront. For several reasons, this is extremely difficult, if not empirically impossible, in the disability transfer case. There are numerous programs providing disability benefits—some public, some private; some means tested, others that limit earnings; some with benefits dependent on severity of disability, others with benefits related to prior earnings; some with federally determined benefit levels and eligibility criteria, others with state-based benefit and eligibility determination; some subsets of programs are integrated, others are not; the application costs of some programs are high, for others the cost is zero; some disability benefits are taxed, others are not. Moreover, multiple benefit recipiency is common. Facing these constraints, an alternative strategy is to model the choice as dependent on total utility, perhaps proxied by total expected income flows, in the various options.

While Parsons's model follows this latter strategy, the numerator in his replacement rate is a seriously deficient indicator of expected utility (income) in the disability option. The imputed value of this benefit is obtained by transforming the observed hourly wage rate in one year (1966) into the payment made to the primary beneficiary through use of the statutory benefit formula. This value totally neglects dependents' benefits that accompany the primary benefit award. Because Parsons imputes this value to all individuals in his sample, the extent of the underestimation of even SSDI benefits is greater than the observed .84 ratio of primary to total benefits for 1980 SSDI recipients, as a smaller percentage of the recipient group are married. Moreover, in addition to SSDI, disabled individuals receive benefits from a plethora of other programs—Workers' Compensation, Veterans Compensation and Disability, Black Lung, Medicare, and a variety of private benefits. Indeed, in 1980, SSDI benefits were less than one-third of total federal disability transfers (excluding Medicare and Medicaid) and a still smaller share of total benefits, public and private. Parsons's *SSB* variable excludes all of these flows. Insofar as Parsons's expected disability income variable captures only a fraction of the total benefits associated with being a recipient, the elasticity implied by its coefficient will be biased upward as an estimate of the impact of SSDI on participation.

---

[3] The numerator is obtained from the 1969 statutory SSDI benefit formula on the basis of the 1966 observed wage rate multiplied by a constant to yield a monthly wage, which is then adjusted by the rate of growth of per capita personal income to yield a covered wage for prior and subsequent years.

Third, as noted above, Parsons's model requires that all observations included in the sample have a reported wage rate. Observations without this measure—primarily those with long-term absences from the work force—are excluded from his analysis, creating a potentially serious sample selectivity problem and a further source of bias in the estimate of the effect of expected *SSB* benefits on participation.

Further, Parsons imputes SSDI benefits to all older males in his sample, a procedure that assumes implicitly that each individual either is eligible for SSDI benefits or can make himself eligible at zero cost. In fact, eligibility for SSDI benefits is conditioned on the demonstration of "total and permanent disability" and requires evidence that no substantial gainful activity (i.e., earnings in excess of $300 per month) is possible[1] As a result of this eligibility assumption, the expected disability transfer income imputed to individuals exceeds their true income expectation in this option, again biasing the coefficient on the replacement rate.[5]

Finally, the vector of control variables in Parsons's [illegible] simonious at best—age, a subsequent mortality-b[illegible] index, the unemployment rate, and an index of [illegible] Where are the variables reflecting family structure [illegible] and needs that are standard control variables in [illegible] and work force participation research?

Although the net effect of the biases from these [illegible] tion and specification procedures is a priori unknown [illegible] on the replacement rate—that variable on which Parsons rests his policy-related assertions—can hardly be considered reliable.

---

[1] Indeed, in 1976, of those who appraised themselves as sufficiently disabled to warrant applying, and thus willing to bear the costs of applying for SSDI benefits (including a 5-month waiting period during which no substantial gainful activity is performed), only 45 percent were actually awarded benefits

[5] A further source of bias is created by the use of the 1966 observed wage as the basis for constructing *SSB* Those individuals who both work in 1966 and receive SSDI benefits will tend to have an observed 1966 wage that is biased downward due to earnings limitations and, as a result, an imputed *SSB* value that is incorrect since the imputed value is not based on earnings prior to receipt of benefits

III

In this section, we test the robustness of Parsons's results, showing the impact on his estimated elasticity of participation response from correcting the misspecifications and erroneous imputation procedures that we have noted. Our purpose is to suggest the magnitude of the biases implicit in Parsons's procedure rather than to argue that the corrected estimate from his basic model is definitive.[6] First we replicate his results using an alternative but richer data base, and then we alter his model, seriatim, so as to measure the effect of each of the corrections. We employ the sample of males 45–62 years of age in the 1978 wave of the Panel Study of Income Dynamics (PSID). The longitudinal nature of the PSID data enables prior actual and imputed wages to be obtained and a disability indicator that reflects the severity of impairment to be constructed. The size of our sample is substantial, though smaller than that of Parsons (741 vs. 3,219). As a result, our *t*-statistics are likely to be somewhat smaller than his.

Our replication of Parsons's model is reported in column 2 of table 1. The disability indicator is an index of the severity of currently reported disability and substitutes for Parsons's index of subsequent mortality. The coefficient estimate on the replacement rate is very close to that of Parsons (.006 vs. .005), as are those for most of the remaining variables. Our replication implies an elasticity of labor force participation of −.021, compared with Parsons's −.030. Our smaller asymptotic *t*-statistic (1.59 vs. 2.48) is due in part to our smaller sample size, and the positive coefficient on *WEL*/*W* is explained by the rapid growth in state welfare payments and the decrease in the number of families with incomes below $3,000 from 1969 to 1977. The correlation between *SSB* and *W* is .61 and between *SSB*/*W* and *W* is .76, indicating a serious multicollinearity problem in Parsons's replacement rate, the crucial variable in his analysis.

The impact of this multicollinearity problem is shown in column 3, where the numerator (*SSB*) and the denominator (*W*) of the replacement rate are entered separately into the model. With this specification, designed to test the direct impact of transfer benefit generosity on labor supply, controlling for the wage, neither the transfer benefit nor the wage variables have an asymptotic *t*-statistic in excess of 0.6. Some of the variation in *LFP* attributed by Parsons to benefit generosity is, in fact, attributable to variation in the wage rate.

In column 4 of the table, we alter Parsons's model in two ways. First, a replacement rate in which the numerator (*SSBD*) includes both the primary and dependent benefits is substituted for *SSB*.[7] Second, control variables reflecting family structure, other household income, and normal occupation are added. These alterations reduce the coefficient on the replacement rate to about one-fourth of its value in column 2 and drop the *t*-statistic to a highly insignificant −0.43. The elasticity measure falls to −.0069.[8]

---

[6] In subsequent research, we have estimated a multistage rational choice model of the effect of all disability transfer income on work status that avoids the estimation problems of the Parsons paper (Haveman and Wolfe, in press).

The effects on Parsons's estimates of his failure to address the selectivity problem are shown in column 5. To correct for potential selectivity bias, a probit equation predicting the presence of the observed wage in 1975 was fit to the entire sample.[9] Using the inverse Mills ratio from this equation along with variables representing age, experience, education, race, and disability status, an OLS equation was fitted to the observations with a reported 1975 wage. Using this equation with reported economic-demographic characteristics, we imputed a 1975 wage to those observations without a reported wage (Heckman 1976). Then we fitted the Parsons model to the larger sample. The impact on the replacement rate is again negative: the coefficient falls to −.0032 and is insignificant. (The asymptotic *t*-statistic is −0.85.) The elasticity falls to −.018.[10]

In column 6, for Parsons's disability transfer variable we substitute one which incorporates an estimated eligibility probability conditional on the individual's health status and other characteristics (*PSSB*).[11] In this specification, the multicollinearity problem is *de minimis* (.03); hence the expected transfer income and wage variables are entered separately. While both the transfer and wage variables have the expected sign, neither has an asymptotic *t*-statistic in excess of 0.9. The elasticity on the disability transfer variable falls to −.0067.

---

[7] *This substitution reduces the multicollinearity problem*
SSBD and *W* is .37

[8] *A specification with dependents benefits added but without* indicates that the additional dependents benefits alone have *coefficient on the replacement rate falls to* −.0026 (−0.86) −.0012

[9] *The variables included in the probit and their asymptotic t-sta* (−2.15); dummy variable for not married and no children = 1 variable for married and no children = 1 (0.87), years of work experience as of 19 (3.55); years of schooling (−0.18); dummy variable for black = 1 (−1.01), age in 1974 times years of schooling (0.34), dummy variables for Protestant = 1 (−1.80), Catholic = 1 (−1.75), and Jewish = 1 (−0.30), and percentage disabled in 1974 (−8.87)

[10] The coefficient, asymptotic *t*-statistic, and elasticity fall to .0015, 0.39, and −.0076, respectively, when the additional control variables are added to the model.

[11] The expected probability of each individual's eligibility for benefits is imputed using the coefficients from a probit equation explaining the acceptance-denial decisions for a group of 45–62-year-old male recent applicants for SSDI benefits included in the 1978 Social Security Administration Survey of the Disabled (SD). Variables available on both the SD and PSID files, including age, education, marital status, race, and disability status, were used. From the estimated coefficients, adjusted to reflect the high proportion of beneficiaries on the SD file, we imputed to each observation on the PSID file an estimate of the probability of being eligible for SSDI benefits, having applied

Finally, in column 7, we combine several of the corrections. The expected benefits/expected wage split is maintained, and the eligibility adjustment, the selectivity correction, and the additional control variables are included. This combination of corrections to the Parsons model suggests a still lower likelihood of a serious negative effect of expected SSDI transfers on the labor force decision. The value of the asymptotic *t*-statistic for the coefficient on the expected benefits variable is a very small −0.48, and the elasticity on the expected benefits variable is −.0056, one-fifth of its value in our replication of Parsons's model.[12]

One remaining issue should be mentioned. Parsons's results are limited by the inadequate subsequent mortality measure of disability status he employs. This index is a weak proxy for work limitations—the primary consideration—since numerous work-related impairments have little relationship to subsequent mortality. As Parsons has pointed out elsewhere (Parsons 1982), no ideal measurement exists in survey data. The most common of those present in survey data are self-reports of disability, and such measures may well be endogenous in a labor supply model (Chirikos and Nestel 1981). While our replication of Parsons's results uses such self-reported information, it is able to measure the severity of the reported disability condition. Moreover, recent studies have indicated that self-reports of health are stable over time, are highly correlated with medical doctor reports, predict future medical evaluation better than the earlier physician assessment, and show no evidence of exaggeration of health problems related to being out of the labor force (Maddox and Douglass 1973; Waldron, Herold, and Dunn 1982). Nevertheless, neither health variable is likely to be anything more than a crude proxy.

IV

Parsons's coefficient estimates are employed in a simulation model that, he finds, closely tracks the increase in the older male nonparticipation rate from .051 to .084 over the 1968–76 period. From this, he concludes that the increasing generosity of SSDI benefits accounts for most of the decrease in participation. In fact, this simulation says little about the causal relationships among these similar time-series patterns. Because of changes in the age composition of the labor force, labor market opportunities for older workers, the incidence of work-related impairments, the level of employment and earnings of spouses, the application of eligibility standards, and the coverage of private pensions, none of which are reflected in Parsons's simulation, his conclusion is unwarranted.

---

[12] As in the col. 6 variant, multicollinearity is not a problem in this specification. If in addition to these corrections to Parsons's model we included the imputed dependent benefits in the expected transfer income variable, the sign of the disability transfer variable reverses, but no statistical significance is indicated.

Indeed, Parsons's own elasticity of $-.03$ indicates that the absolute impact of, say, a 10 percent increase in benefits (with the wage rate held constant) is only about 45,000 older males, 0.3 percent of an older male work force of nearly 14 million. Reliance on this elasticity value to suggest that "the decline in labor force participation is the result of increasingly attractive alternatives to work" (p. 131) is misleading at best. Indeed, the corrections of his specification and primary variables presented in this paper suggest that SSDI benefits have only a small adverse labor supply effect.[13]

References

Chirikos, Thomas N., and Nestel, Gilbert. "Impairment and Labor Market Outcomes: A Cross-sectional and Longitudinal Analysis." In *Work and Retirement: A Longitudinal Study of Men*, edited by Herbert S. Parnes. Cambridge, Mass.: M.I.T. Press, 1981.

Haveman, Robert H., and Wolfe, Barbara L. "Disability Transfers and Early Retirement: A Causal Relationship?" *J. Public Econ.* (in press).

Heckman, James J. "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals Econ. and Soc. Measurement* 5 (Fall 1976): 475–92.

Maddox, George L., and Douglass, Elizabeth B. "Self-Assessment of Health: A Longitudinal Study of Elderly Subjects." *J. Health and Soc. Behavior* 14 (March 1973): 87–93.

Parnes, Herbert S.; Eggs, Karl; Kohen, Andrew I., and
The Pre-Retirement Years: A Longitudinal Study of the Lab
Men. Vol. 2. Washington: Government Printing Offi

Parsons, Donald O. "The Decline in Male Labor Force
88 (February 1980): 117–34. (*a*)

———. "Racial Trends in Male Labor Force Participa
cember 1980): 911–20. (*b*)

———. "The Male Labour Force Participation Decision
Health, and Economic Incentives." *Economica* 49 (Februar

Waldron, Ingrid; Herold, Joan; and Dunn, Dennis. "How Valid Are Self-Report Measures for Evaluating Relationships between Women's Health and Labor Force Participation?" *Women and Health* 7 (Summer 1982): 53–66.

---

[13] In Haveman and Wolfe (in press), we find the coefficient on expected disability transfer income to be statistically significant, though small. The elasticity of labor force participation with respect to expected disability transfers ranged, at the mean, from $-.0003$ to $-.0005$. This estimated response is lower than any of those reported in table 1. The estimated elasticity is nonlinear, however, and the response is greatest for those with low earnings capacity.

Disability Insurance and Male Labor Force Participation: A Response to Haveman and Wolfe

Donald O. Parsons

Ohio State University

The percentage of adult males 45–54 years of age who are out of the labor force rose from 4.5 percent to 8.7 percent between 1947 and 1982. A great deal is known about this unprecedented rise in nonparticipation in the labor force of middle-aged males, much more than would be apparent from reading Haveman and Wolfe's comment. Substantial evidence has accumulated that the labor force participation of middle-aged males is sensitive both to the characteristics of the social security disability system and to labor market alternatives.[1]

Haveman and Wolfe's empirical results are an anomaly in this body of literature. Their generally insignificant empirical results are apparently not the result of the specification issues they raise but rather of their unfortunate choice of data set. An earlier study by Leonard (1979), for example, was almost identical in design to the Haveman and Wolfe analysis yet resulted in parameter estimates and conclusions remarkably similar to my own. Haveman and Wolfe's sample unfortunately is one-half the size of Leonard's sample and one-fifth the size of mine and apparently lacks detail in the crucial health measure.

The issues Haveman and Wolfe raise are issues of variable measurement. They do not take issue with the fundamental choice structure underlying my empirical model, an expected utility framework discussed in Parsons (1980*a*). Since the disability system requires that the individual be incapable of "substantial gainful activity," potential applicants face a labor force participation decision, not a marginal hours variation decision. A standard expected utility analysis leads to a probit model for the probability of labor force participation (LFP)

---

[1] See particularly Lando (1979), Lando, Coate, and Kraus (1979), Leonard (1979), Parsons (1980*b*, 1982*a*, 1982*b*, 1982*c*), and Marvel (1982). A more detailed summary of recent social security disability estimates is available on request from the author.

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9203-0008$01.50

of the form $LFP = F(I)$, where F is the cumulative normal function and $I = I(W, UR, B, \phi, H, Z)$ with $W \equiv$ wage rate, $UR \equiv$ unemployment rate, $B \equiv$ benefit level of disability insurance, $\phi \equiv$ probability that a disability claim will be accepted, $H \equiv$ index of health condition, and $Z \equiv$ vector of labor supply factors. The form of the index function varies, of course, with the form of the individual's utility function; a Cobb-Douglas form, for example, implies a log-linear index function with the log of the replacement rate (B/E) as one of the arguments. Generally researchers have not assumed a detailed knowledge of the individual's utility function and have instead estimated a linear approximation to the index function.

Suggested Modifications in Variable Measurement

Haveman and Wolfe suggest refinements in the measurement of the individual's wages, benefits, and likelihood of a successful disability claim. Their concern about my wage measure is, I suspect, the result of their misunderstanding of one of my related papers on this subject (Parsons 1980*b*), which they cite but perhaps have not read carefully. Obviously there is a general measurement problem with the wage variable since individuals who are not working have no observable wage. In the paper on which Haveman and Wolfe comment, I use (3-year) lagged wages to reduce the potential selectivity problem, although truly long-term nonparticipants would still be underrepresented. In a related paper (Parsons 1980*b*), however, I use an instrumental variable approach much like the one they
find that the estimated nonparticipation elasticity
social security replacement rate is larger, not s
mate based on lagged wages. Leonard used lagge
but had lifetime earnings records from social se
did not have a selectivity problem; virtually all
point.

The question of the disability award probability
The result of an elaborate adjudication process, award
unobservable for most individuals in this sample. Since the primary criterion is health condition, I estimated and reported a model that included an interaction term for health and the replacement rate, this alternative specification did not affect my conclusions. Both Leonard and Haveman and Wolfe use a more elaborate instrument for the award probability, based on social security records of acceptance and denials of applications. If one assumes that those who apply for disability benefits have the same unobserved characteristics as those who do not, this is appropriate. Otherwise a sample selectivity problem arises.

Finally, Haveman and Wolfe make a number of suggestions for the measurement of benefits. They make what appears to be an incorrect statement that the possibility that I measure only a fraction of the returns to nonparticipation will inflate my elasticity estimate.[2] Of more substance, they argue that benefits should be entered separately from wage rates, rather than as the replacement rate. Clearly since benefits are computed by social security based on past earnings, a multicollinearity problem exists here. Leonard avoided this problem by entering expected benefits (award likelihood times benefits), not benefits, in the model, and he entered the two variables separately. Haveman and Wolfe follow a similar strategy.

Leonard and Haveman and Wolfe also include potential dependent benefits in their benefit measure. A spouse 62 years of age or older or, more frequently in this age group, dependents 18 years of age or less could receive benefits based on the male's disability claim; in recent years such benefits have comprised approximately 17 percent of DI trust fund disbursements. This suggestion is superficially appealing but raises substantial problems of estimation. Considerable evidence exists that family structure and therefore dependent benefits themselves depend on the disability status of the male head (Franklin 1977) and in a subtle way on the male's labor force status as well (Marvel 1979). More careful consideration of family effects is certainly warranted; the simple inclusion of dependent benefits in the benefit measure may not, however, lead to superior estimates. The relatively small magnitude of these payments suggests, moreover, that the estimates may not be sensitive to the precise handling of this concept.

The Consequences

Although the suggested refinements of the variable measurements would appear to be modest, the impact on the estimates of the model is ultimately an empirical question. What, then, is the consequence of incorporating Haveman and Wolfe's proposed modifications? Apparently very little, if one accepts Leonard's estimates. Leonard reports an elasticity of nonparticipation with respect to expected social security benefits of 0.35, roughly comparable to my own estimates of 0.49 (based on the lagged wage approach) and 0.93 (based on the instrumental variable approach).[3] The independently estimated wage coefficient is roughly equal in magnitude and opposite in sign, suggesting that the coefficient constraint implicit in the use of the replacement rate may be of no substantial consequence. Even the somewhat lower estimated elasticity may be the consequence of Leonard's use of a self-reported health measure. In recent papers, Parsons (1982*a*) and Anderson and Burkhauser (1983) demonstrate that self-reported health measures are determined jointly with labor force participation for individuals in this age-demographic group and that the inclusion of such measures biases toward zero the estimated incentive effects of social security benefits. Leonard recognized this problem and used the responses to quite detailed health condition questions to eliminate the crudest sort of response bias, but some problem may remain.

---

[2] Haveman and Wolfe remark, "Insofar as Parsons's expected disability income variable captures only a fraction of the total benefits associated with being a recipient, the elasticity implied by its coefficient will be biased upward as an estimate of the impact of SSDI on participation."

[3] The elasticity measures reported here differ from those reported in the original. The incorrect labor force participation mean was used in the computation of the participation elasticity in the elasticity estimate reported in Parsons (1980*a*). A minor error in the wage instrument of Parsons (1980*b*) has also been corrected.

Haveman and Wolfe, of course, also estimated such a model, although on a much smaller data set. Even their "replication" of my original model resulted in no economic variable significant at customary levels, a finding quite at variance with my work and Leonard's and also with that of others who have undertaken research in this area.[4] Even the unemployment rate, which has been found to affect labor force participation and disability application rates in a wide range of cross-section and time-series analyses, is insignificant in Haveman and Wolfe's model.

The relatively small size of their data set would not seem sufficient to explain these differences. Haveman and Wolfe's anomalous estimates are more likely the consequence of the health measure they introduce into the model. The measure, "an index of the severity of currently reported disability," is never defined precisely
not possible to comment on it in detail. Th
justification) problem with the use of a self-repo
in labor supply equations has been recognized f
man and Wolfe justify their use of the self-report
ence to two papers, one a tabular analysis of t
behavior of older women (Waldron, Herold, an
the other no more than a demonstration that self-rep
frequently accord with clinical measures (Maddox and Douglass 1973). At the same time they ignore recent evidence reported by me and by Anderson and Burkhauser that the use of self-reported health measures introduces serious problems into the estimation of labor supply models for the demographic group in question, older males. In any case, the absence of any significant economic effects in their "replication" model makes the subsequent sensitivity analysis a curious exercise, from which one can extract little of interest.

---

[4] See the references cited in n. 1, all of which report strong and significant disability program and market effects in labor force participation or social security recipiency models.

not, however, divert attention from the fact that substantial progress is being made toward understanding the economic determinants of middle-aged male labor force participation and disability recipiency. This demographic group is sensitive to the economic reward structure of the disability program and the market.

APPENDIX

DATA SERIES FOR FIGURE 1

Year (1)	Labor Force Nonparticipation Rate of Males 45–54 (%) (2)	Social Security Disability Rate of Males 45–54 (%)* (3)
1947	4.5	.00
1948	4.2	.00
1949	4.4	.00
1950	4.2	.00
1951	4.1	.00
1952	3.8	.00
1953	3.5	.00
1954	3.5	.00
1955	3.5	.00
1956	3.5	.00
1957	3.7	.23
1958	3.7	.36
1959	4.0	.51
1960	4.3	.76
1961	4.4	1.13
1962	4.3	1.54
1963	4.3	1.54
1964	4.3	1.63
1965	4.4	1.79
1966	4.7	1.96
1967	4.8	2.11
1968	5.1	2.25
1969	5.4	2.37
1970	5.8	2.51
1971	6.1	2.75
1972	6.8	3.03
1973	7.0	3.27
1974	7.8	3.50
1975	7.9	3.84
1976	8.4	4.03
1977	8.8	4.23
1978	8.7	4.25
1979	8.6	4.17
1980	8.8	N.A.
1981	8.6	N.A.
1982	8.7	N.A.

SOURCES.—Participation Rate and Population by Age, *Handbook of Labor Statistics 1980*, *Employment and Training Report of the President 1982*, Disabled by Age *Social Security Bulletin*, *Annual Statistical Supplement*, 1981.

* The ratio of male disability recipients 45–54 years of age to the number of males 45–54 years of age.

References

Anderson, Kathryn H., and Burkhauser, Richard V. "The Effect of Actual Mortality Experience within a Retirement Decision Model." Working Paper no. 83-W08, Vanderbilt Univ., 1983.

Butler, Richard, and Heckman, James J. "The Government's Impact on the Labor Market Status of Black Males: A Critical Review." 1977 Industrial Relations Research Association Series, *Equal Rights and Industrial Relations* Pp. 235–80.

Franklin, Paula A. "Impact of Disability on the Family Structure." *Social Security Bull.* 40 (May 1977): 3–18.

Gastwirth, Joseph L. "On the Decline of Male Labor Force Participation." *Monthly Labor Rev.* 95 (October 1972): 44–46.

Lando, Mordechai. "Prevalence of Work Disability by State, 1976." *Social Security Bull.* 42 (May 1979): 41–44.

Lando, Mordechai E.; Coate, Malcolm B.; and Kraus, Ruth. "Disability Benefit Applications and the Economy." *Social Security Bull.* 42 (October 1979): 3–10.

Leonard, Jonathan S. "The Social Security Disability Program and Labor Force Participation." Working Paper no. 392, N.B.E.R., August 1979.

Maddox, George L., and Douglass, Elizabeth B. "Self-Assessment of Health: A Longitudinal Study of Elderly Subjects." *J. Health and Soc Behavior* 14 (March 1973): 87–93.

Marvel, Howard P. "Health and the Economics of the Family." Mimeographed. Columbus: Ohio State Univ, 1979.

———. "An Economic Analysis of the Operation of Social Security Disability Insurance." *J. Human Resources* 17 (Summer 1982): 393–412.

Parsons, Donald O. "The Decline in Male Labor Force Participation." *J.P.E.* 88 (February 1980): 117–34. (*a*)

———. "Racial Trends in Male Labor Force Participation." *A.E.R.* 70 (December 1980): 911–20. (*b*)

———. "The Male Labour Force Participation Decisio Health, and Economic Incentives." *Economica* 49 (Feb (*a*)

———. "A Time Series Analysis of Male Labor Force Par Security Disability Recipiency." Mimeographed. C Univ., 1982. (*b*)

———. "Male Labor Force Participation and Disability Ke State Analysis." Mimeographed. Columbus, Ohio State U

Siskind, Frederic B. "Labor Force Participation of Men, 25– *Monthly Labor Rev* 98 (July 1975): 40–42.

Swisher, Idella G. "The Disabled and the Decline in Men's Labor Force Participation." *Monthly Labor Rev.* 96 (November 1973) 53.

U.S. Bureau of the Census "Occupations at the Twelfth Census." Special Report of *The 1900 Census of Population.* Washington Government Printing Office, 1904.

Waldron, Ingrid; Herold, Joan; and Dunn, Dennis. "How Valid Are Self-Report Measures for Evaluating Relationships between Women's Health and Labor Force Participation?" *Women and Health* 7 (Summer 1982). 53–66.

Book Reviews

Import Competition and Response. Edited by JAGDISH N. BHAGWATI. Chicago: University of Chicago Press (for National Bureau of Economic Research), 1982. Pp. ix+410. $32.50 (cloth); $15.00 (paper).

The rapid incursion of imports into markets long dominated by domestic producers has raised urgent policy questions in the United States and Europe. Whether and how national governments should respond to changing international competitiveness is now a central issue for public debate in the major industrial nations. Yet while policymakers have been forced by circumstances to improvise means for coping with the rapid growth of imports, especially imports from Japan and the newly industrializing countries (NICs), theorists have tended to focus their energies elsewhere.

Import Competition and Response may be the first major collection of papers attempting to provide some theoretical guidance for a policy response to rapidly changing conditions of trade. The volume, which results from a May 1980 conference of the National Bureau of Economic Research, contains contributions by a number of well-known theorists. The purpose of the conference was to develop theoretical perspectives on the alternative policy responses to import competition. Not surprisingly, the theoretical papers largely represent new twists on themes already introduced in other places. The result, as the editor evaluates it in the preface, is a volume addressing a number of theoretical issues that are "fairly novel and which also bear on policy questions, in many cases fairly directly." Thus, those seeking immediate policy relevance are apt to be disappointed, and in any case the contributions are too technical to exert a direct influence in policy circles. But for economists concerned with narrowing the conspicuous gap between theoretical prescriptions and the actual policy process, the book offers ample food for thought.

The first four papers analyze the role of policy in the adjustment of an economy to a change in the international terms of trade. Michael Bruno's model of macroeconomic adjustment with wage-price rigidity demonstrates that the effect of import competition cannot be treated without considering the prior state of the economy. Thus, according to Bruno, increased import competition from NICs during the 1970s had a magnified consequence for macroeconomic performance because it occurred while the industrialized nations were still reeling from the effects of oil shocks.

Wage stickiness is also assumed by J. Peter Neary in a new variant of the two-sector model with sector-specific capital. Neary shows that with wages sticky both upward and downward, the economy's adjustment path may be

[*Journal of Political Economy*, 1984, vol. 92, no. 3]
© 1984 by The University of Chicago. All rights reserved.

characterized by a temporary fall in national income evaluated at world prices—a possibility he calls "immiserizing reallocation" of productive factors. Neary assumes also that the reallocation of capital between sectors depends only on *current* rates of return, even though those rates are changing as adjustment proceeds. It is apparently the interaction of this myopia with sluggish wages that accounts for the temporary fall in income.

Michael Mussa's model also highlights the constraints on the speed of capital adjustment. However, Mussa does this through an explicit cost in terms of labor of moving capital between sectors, rather than simply assuming, as Neary does, a finite rate of adjustment that depends on relative rates of return. The focus of Mussa's analysis is the potential role for government in the adjustment process. Mussa shows that, in the absence of distortions, private maximizing behavior produces socially optimal, albeit gradual, adjustment. However, the mere presence of distortions is not sufficient to ensure that intervention in the adjustment process will improve matters.

What is striking in these three papers is the almost negligible attention to issues of distribution. As Alasdair Smith suggests in his comment on Mussa, it is debatable whether a model in which income distributional issues are absent can do justice to real-world motivations for intervention in the adjustment process. Yet arguments for real-world policies often do claim precisely that intervention is required for reasons of economic efficiency; indeed, the role of these policies in redistributing income is usually soft-pedaled by their sponsors. One would thus like to be sure that the policy-minded reader comes away not only with Neary's message, that in a world of distortions a case for adjustment assistance on purely efficiency grounds can be constructed, but also with Mussa's more complicated conclusions, which point to the limited value in achieving *efficiency* of policies commonly advocated to facilitate adjustment.

In contrast to the other three, income distribution is the centerpiece of Peter Diamond's analysis of trade adjustment assistance. Diamond's model is based on an industry such as footwear, in which a fall in the world price has produced wages below those elsewhere in the economy. In this context he weighs the roles of subsidies to workers remaining in the import-impacted industry, which increase social welfare by redistributing inco [illegible] sidies to moving costs, which promote appropriate ad [illegible] world price. However, it is interesting to juxtapose this [illegible] Richardson's evaluation of actual trade adjustment ass [illegible] States under the Trade Act of 1974, in the "empirical [illegible] end of the book. There Richardson points out that the [illegible] after 1974 went to workers earning above-average inco [illegible] as autos. By Diamond's logic, workers in the auto indus [illegible] taxed, not subsidized. Richardson also points out that a [illegible] workers eventually returned to well-paid employment in the sa [illegible]

A second section of the book consists of five papers dealing with "political economy" issues. These papers identify the incentives for lobbying and political action implied by actual or potential changes in trading circumstances. In a paper rich in testable hypotheses and links to current policy issues, Jagdish N. Bhagwati contrasts the responses of two types of industries undergoing a change in international competitiveness: senescent low-technology production and dynamic, technologically progressive industries. Bhagwati's analysis incorporates the role not only of tariffs but also of international migration and direct investment.

Paul Krugman uses a model of trade in differentiated products to show that bilateral trade liberalization may be mutually beneficial for industries in which product differentiation rather than traditional comparative advantage forms the basis for trade. Readers interested in the general subject of modeling product differentiation may profit also from the comments on Krugman's paper by Kelvin Lancaster, Michael Mussa, and John Chipman, all of whom deal with the technical properties of such models.

Ronald Findlay and Stanislaw Wellisz provide a general equilibrium model in which the tariff rate is determined endogenously as a consequence of rational resource allocation to lobbying activities by interested parties. As Richard A. Brecher notes in a comment, one restriction of the model's applicability is that it assumes a minimum value of zero for the tariff. This rules out the empirically interesting case of export interests' pressing for subsidies. In a somewhat similar vein, Robert C. Feenstra and Jagdish N. Bhagwati explore the "efficient" use of direct compensation, financed from tariff revenues, to induce special interests to accept a lower tariff rate. The final paper in this section, by Robert E. Baldwin, summarizes and extends the theory of trade policy determination and reviews the results of recent empirical attempts to account for industry differences in access to protection.

The nine theoretical papers offer a rich and varied menu of approaches to the problems of import competition (a short introductory chapter by Bhagwati weaves them all into a coherent framework). What might have appeared here but did not is best suggested by the three papers that close the book. Although by no means giving a full picture of the current policy environment, the "case studies" certainly point to a significant disjunction between the theories and the processes they are intended to illuminate.

Ronald Dore's account of a British town once dominated by textile production suggests that, even in the absence of government policies, adjustment for both firms and workers was decidedly less painful than usually assumed. As Richardson notes in his comment on Dore, "it does only a little violence to this excellent paper to characterize it as documenting a history of gradually successful private adjustment and questionably successful official efforts to ease that adjustment." Richardson's own survey of trade adjustment assistance in the United States, based on survey data in which recipients of standard unemployment insurance served as a control group, contains a variety of unexpected findings that would be highly relevant to the design of a new program to facilitate adjustment However, it is especially striking in this context that the one major instance in which trade adjustment assistance has actually been implemented produced a set of policies bearing little resemblance to those analyzed in the theoretical section of the book.

The final paper, an assessment of protectionist policies (especially toward developing countries) in the European Community, takes a political economy approach unlike that of the theorists, emphasizing in particular the critical role of government structure. Eric Verreydt and Jean Waelbroeck give on the whole an eminently sensible account, although the theoretically minded reader may puzzle over an argument attributing the strength of protectionist pressures in steel to inelastic industry supply juxtaposed with one just a page or two earlier that highly *elastic* supply is responsible for the success of textile and clothing interests in obtaining protection. Notable also in the paper is the prominent role of protection through voluntary export restraints (VERs), a tool never mentioned in the theoretical portions of the book. As some recent literature has shown, the effects of VERs differ in important ways from those of tariffs or conventional import quotas, particularly with regard to distribution of costs and benefits and implications for competitive behavior.

Several further oddities leap to the eye of the reader seeking policy relevance. The theoretical papers keep to the usual tradition of viewing unexploited national monopoly power in world markets as a distortion. In other words, welfare analysis is from the national rather than global point of view, and the issue of foreign retaliation is not raised. Likewise, the analyses of political economy largely ignore that the main lobbying for freer trade in the United States and the European Community has come from export-oriented industry, from those whose business is trade itself, and from *industrial* consumers of imported intermediate inputs, the last noted by Verreydt and Waelbroeck. Recent political history gives little support to a key role for protests from ultimate consumers. Increased protection of the automobile industry in the United States may prove an exception, but probably as much because of populist reaction to the high average wage in the industry as the direct pocketbook impact at the showroom.

Finally, the theoretical chapters contain almost no mention (Bhagwati's comes close to an exception) that it no longer makes sense—if indeed it ever did—to think of "international" policy separately from "domestic" policy. This point does come out obliquely in Richardson's paper and strongly in Verreydt and Waelbroeck. An implication is that the empirical studies surveyed by Baldwin need to be interpreted with great caution, since tariffs represent only one of many possibilities for granting favorable treatment to a politically potent industry.

Despite an unevenness in quality and approach that is inevitable in a conference volume, this book offers valuable and at times even exciting reading for the professional economist. Yet, as the comments above suggest, it represents only a beginning in a difficult endeavor. Given the stated aim of providing "respectable theoretical work that . . . enables the empirical analyst to examine the phenomenon insightfully," the theoretical papers
untainted by even casual reference to current policy iss
ment on Verreydt and Waelbroeck by Gene M. Grossi
ical contribution in the entire volume that fully de
respectable theory linked to concrete suggestions for
tion. Stanislaw Wellisz's summary assessment of Bald
economy approaches in fact applies equally well to the
be humbled by how little we have accomplished, but
need not despair that nothing is left to conquer."

[illegible]

University of Wisconsin—Madison

Inflation: Causes and Effects. Edited by ROBERT E. HALL.
Chicago: University of Chicago Press (for National Bureau of Economic Research), 1982. Pp. 290. $30.00 (cloth); $10.95 (paper)

The cover flap of this book proclaims: "this volume presents the latest thoughts of a brilliant group of young economists on one of the most persistent economic problems facing the United States and the world, inflation." This group of economists is brilliant, and inflation is an inveterate problem, but somewhat disappointingly some of the papers in this volume fall shy of the high standard of work that the profession has come to take for granted from these individuals.

Perhaps the most interesting paper in the book is Thomas Sargent's study of the ends of four hyperinflations. This paper is undoubtedly going to be viewed as provocative by many, but it is also likely to be important reading for the macroeconomist. The central theme of the paper is that monetary and fiscal policy cannot be analyzed in isolation from each other. The four hyperinflations studied by Sargent—interwar Austria, Germany, Hungary, and Poland—all were characterized by economies with near fiscally bankrupt governments. That is, the present value of these governments' expenditures and obligations far outstretched the present value of their taxation capabilities. These hyperinflations were halted only when credible fiscal reforms were undertaken. The amazing fact, which Sargent documents, is that inflation stopped almost overnight when these fiscal reforms were enacted, even though in all cases monetary reforms did not take place until some later dates, so that at the time of the fiscal reforms these countries' money supplies were still growing at astronomical rates. Sargent notes that because the switches in these governments' fiscal regimes were largely anticipated, as well as credible, the dramatic reductions in these countries' inflation rates were not associated with massive rises in unemployment, as many people might expect.

Taking the opposite view, Robert Gordon claims that reductions in the inflation rate are likely to be associated with heavy costs in reduced real output—and presumably in higher unemployment. To substantiate this claim, he plots the relationship between the inflation and "adjusted nominal GNP growth"—defined as the difference between nominal GNP growth and the trend growth of real income—for the United States and eight other countries. On a number of these graphs, it can be seen that movements in adjusted nominal GNP growth were not associated with one-to-one movements in the inflation rate. By definition, any slippage between these two numbers must be made up by movements in the detrended growth rate of real GNP. Gordon asserts that this is evidence that anti-inflationary stabilization policy has a high cost since changes in adjusted nominal GNP growth were often associated with significant movements in the detrended growth rate of real GNP.

Now, the reader can clearly see that the pattern of association between the inflation rate and adjusted nominal GNP growth is not stable across either time periods or countries—in fact, Gordon acknowledges this. Furthermore, as Lucas (1976) has discussed, these types of statistical relationships are unlikely to be invariant with respect to government stabilization policy. Unfortunately, Gordon makes no attempt to break down movements in his adjusted nominal GNP growth variable so as to isolate the effects of government stabilization policy while controlling for the public's expectations of these policies. Of course, one way to do this would be to dichotomize movements in the money supply into anticipated and unanticipated components as Barro (1977) has done. As is well known, Barro's work weighs strongly against the hypothesis that anticipated reductions in the money supply lead to substantial drops in real output. In fact, many of Gordon's graphs seem to show that changes in his nominal income variable are mainly absorbed by movements in the price level. Gordon unconvincingly argues that these results vis-à-vis the others are statistical anomalies due to inappropriate detrending or other factors.

It is hard to see how these graphs can be used for any purpose other than providing a pictorial history between the two plotted variables. Unfortunately, this kind of casual empiricism is not without cost. Another author in the book, Douglas Hibbs, cites evidence such as Gordon's as justification that a policy of wage and price controls cum monetary restraint might be considered a painless method of curing inflation. This is just the kind of policy question that this type of analysis cannot be used to address scientifically.

How to control inflation is an important question. On this topic the book contains two papers on monetary standards: one by Robert Barro and the other by Robert Hall. Barro starts off by noting that for postwar U.S. history, the relationship between monetary growth and inflation over reasonably long periods of time is fairly close. Over shorter periods of time, this association is not so strong. He attributes this to swings in velocity. He notes that in recent times there have been large fluctuations in nominal interest rates probably due to changes in inflationary expectations. By his estimate, a 1 percent increase in nominal interest rates can cause as much as a 4 percent increase in the measured annual inflation rate due to the once-and-for-all increase in the price level induced by the reduction in the demand for money. Barro argues that the volatility in inflationary expectations is the consequence of the lack of a nominal anchor for the U.S. monetary system. Such an anchor could be provided by either a commodity standard or a monetary constitution that restricted the rate of monetary expansion. Finally, he notes that whatever system is adopted, it must provide the necessary incentives to ensure government compliance with the set-out rules. This is important and a nontrivial problem.

Hall also comments on this last point. He states that there may be no reason on a priori grounds to favor a commodity standard over the present fiat money system. This is because under the former system it may be just as easy to change the price of the commodity unit as it is to change the money supply under the latter system. The same pressures that exist to inflate the economy under a fiat money standard by increasing money supply may still exist to inflate the economy under a commodity standard by increasing the currency price of the commodity unit. Also, obviously the choice of commod [illegible] under a commodity standard is an important question. [illegible] commodity bundle chosen should be one whose value [illegible] line with the general cost of living. With this in mind, he [illegible] ity bundle consisting of ammonium nitrate, copper, alu [illegible]

A postmortem of U.S. inflation in the 1970s is underta [illegible] His analysis consists of dissecting the rate of change in th [illegible] change of its various components. This approach acco [illegible] with the layman's perception of inflation, that it is ca [illegible] harvest failures, or rising mortgage rates. While defini [illegible] taxonomy of the CPI is not likely to form the basis of a m [illegible] inflation. The supply and demand for oil provide an expl [illegible] relative price and not for either the nominal price of oil [illegible] level. Furthermore, any movement in the overall CPI is [illegible] pattern of change in relative prices across industries, [illegible] minds us of in his contribution to the volume. This is not to say that [illegible] specific sectors of the economy cannot affect inflation. They can. A large change in the world price of oil could have a significant impact on the current and future levels of a nation's real output, and consequently on its demand for money, and thus on the time path of its general price level. By undertaking an anatomy of the CPI, Blinder tends to obscure the important point that principal movements in the CPI come about from the interplay between money demand and money supply.

A central question concerning inflation is, What are its welfare costs? Fischer (1981) suggests that there may be several types of costs associated with inflation, such as increased transaction costs of exchange (the traditional money triangle), the misallocative effects on agents' savings and investment decisions of tax systems that are nonneutral with respect to inflation, the distortions imposed by nominal interest rate ceilings, and the costs connected with the increased uncertainty (and relative price variability) that seem to be connected with inflation, inter alia. Attaching costs, arising via these various channels due to inflation, is obviously an exercise in modern welfare economics.

One way to approach this problem would be to construct a small-scale, dynamic, choice-theoretic general equilibrium model and examine the impact on an economy's welfare that changes in the time-series properties of its money supply have through these various channels. For instance, within the construct of a cash-in-advance economy, Stockman (1981) illustrates how an increase in an economy's (steady-state) rate of inflation leads to a decrease in savings, investment, and welfare due to increased transaction costs of exchange. Aschauer (1980) in a similar model finds that high rates of inflation are associated with reduced labor supply, higher unemployment, and lower welfare. Furthermore, it seems that by judiciously picking the functional forms and key parameter values to be used in such small-scale dynamic general equilibrium models, quantitative estimates of inflation's welfare costs could be obtained. Such a modeling strategy would be in the spirit of Kydland and Prescott (1982).

Several papers in the volume do give a loose idea of some of the likely costs associated with inflation, but none of them do it in the formal (and perhaps idealistic) manner suggested above. For instance, Martin Feldstein outlines how the interaction between inflation and the current U.S. tax system has had powerful distortionary effects on the realized real rates of return facing savers and investors. The implications of this for investment in residential and nonresidential capital are discussed. Stanley Fischer notes that the costs of inflation depend on the institutional structure of the economy. Now, one would expect the institutional structure of an economy to be somewhat endogenous, reacting through time to changing economic circumstances. The author documents how certain institutions in the U.S. economy have adapted so as to make the recent inflationary era easier to live with. The phasing-out of interest-rate ceilings by 1986 due to the Depository Institutions Deregulation and Monetary Control Act is one example among many. As Fischer notes, however, there are still major instances of nonadaptations in the American economy, such as the government's failure to index the tax system. Generally, though, the long-run costs of inflation are likely to be significantly less than the short-run costs of inflation because of the endogeneity in the economy's underlying structure.

Another paper that addresses the impact of inflation on the economy is by Dennis Carlton, who argues that inflation has had a disruptive impact on the organization of markets. He believes that as inflation rises, so do the transaction costs of doing many types of businesses. When the inflation rate rises, firms will tend to substitute away from those business activities whose relative transaction costs have increased. For instance, he predicts that a rising inflation rate will lead to firms switching away from customized products—either internally produced or purchased from outside—to standardized ones instead. Additionally, where customized inputs are still used, firms will tend to produce these inputs themselves rather than purchase them from an outside party. Thus, on the latter account, there would be an increase in vertical integration. These types of substitution effects induced by inflation may be very important.

Inflation increases the uncertainty of doing business. Carlton notes that this has led to fixed-price contracts of shorter duration between firms. If inflation does create great uncertainty in the economic system, there is likely to be a greater discrepancy in people's beliefs about the course of future economic events. This creates an incentive for people who think they hold superior knowledge to act on their beliefs by participating in futures markets. Carlton notes that both the amount of new futures contracts introduced and the volume of trading in futures markets have increased with the inflation rate.

These are other examples of how the institutional structure of an economy adjusts so as to ameliorate the costs of inflation. The paper by Carlton is interesting because it illustrates how inflation may have significant costs by disrupting the allocative efficiency of the economic system. These types of costs, while potentially pervasive, may be of so subtle a nature that their channels of effect have been too elusive for the economist to delineate precisely. Finally, it must be said that while many of this author's hypotheses seem to have only the status of educated conjectures, they certainly are ones that merit further theoretical and empirical research.

The book also contains articles by four other individuals. Room is left here for only a capsule summary of these papers. Jeremy Bulow discusses the effects that inflation has had on American private pension plans. He finds that these plans have been amassing huge surpluses. The impact of inflation on measured American corporate profits is analyzed by Jeremy Bulow and John Shoven. They find that, when measured appropriately, corporate profits are higher in the recent inflationary era than is normally supposed. Douglas Hibbs statistically examines the interaction between the state of the macroeconomy and public opinion about inflation and unemployment. Finally, Jacob Frenkel addresses several open-economy aspects of the recent U.S. inflation.

While most of the work in this book is of good quality, some of it is not of the unusually high caliber that the profession has come to exp
talented group. Because of this, those interested in the
this volume would be well advised to examine it in the
to place their orders.

University of Western Ontario

References

Aschauer, David. "Welfare, Unemployment and Anticipated Inflation in a Pure Monetary Economy." Unpublished paper. Rochester, N.Y.: Univ. Rochester, March 1980.

Barro, Robert J. "Unanticipated Money Growth and Unemployment in the United States." *A.E.R.* 67 (March 1977): 101–15

Fischer, Stanley. "Towards an Understanding of the Costs of Inflation." In *The Costs and Consequences of Inflation*, edited by Karl Brunner and Allan H Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 15. Amsterdam: North-Holland, 1981.

Kydland, Finn E., and Prescott, Edward C. "Time to Build and Aggregate Fluctuations." *Econometrica* 50 (November 1982): 1345–70.

Lucas, Robert E., Jr. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets*, edited by Karl Brunner and Allan H. Meltzer. Carnegie-Rochester Conference Series on Public Policy, vol. 1. Amsterdam: North-Holland, 1976.
Stockman, Alan C. "Anticipated Inflation and the Capital Stock in a Cash-in-Advance Economy." *J. Monetary Econ.* 8 (November 1981): 387–93.

The Political Economy of Growth. Edited by **Dennis C. Mueller.**
New Haven, Conn., and London: Yale University Press, 1983. Pp. 285. $23.50.

As stagflation became an increasing concern in the 1970s, many economists investigated possible economic "causes" of the slowdown in productivity growth, including the relationship between inflation and slower growth, changing demographic structures, and tax-induced disincentives to save. Simultaneously, Mancur Olson was developing an intriguing and more political-economic explanation of the phenomenon. Following on his earlier seminal work on the logic of collective action, Olson focused on the determinants of economic policies as a factor in growth. In his view, in societies with stable sociopolitical structures, interest groups tend to spring up and become stronger over time. The more interest groups defend their interests, the more growth retardation there is, for these groups seek to prevent change that might come about at the expense of their particular (short-term) interests.

The Mueller volume consists of a series of papers from a conference of which the Olson hypothesis was the theme. The first paper in the volume is Olson's. Although it is less persuasive than his book (Olson 1982) and predates the book by about 4 years, it nonetheless sets forth the main arguments of the interest group hypothesis: different countries have experienced different growth rates at different times, so differences cannot be attributed to different social or cultural patterns; although there are costs to forming special interest groups and the benefits of collective action have some aspects of a public good, special interest groups are likely to emerge in homogeneous groups and to be based on relatively special, as contrasted with mass, interests; it requires time for special interest groups to organize and emerge as effective. Hence, "democratic societies will accumulate more special-interest organizations as time goes on. Collective action is difficult and problematic, especially for large groups, and will occur if at all only under favorable conditions. The longer a society has stable freedom of organization, the greater the likelihood that any given group with a common interest will find conditions in which it can work out selective incentives [for individuals to join in collective action]" (pp. 16–17).

The next step in the argument is to posit that most efficiency-increasing policy changes benefit large numbers of people and impose costs on a relatively smaller group. Because a smaller group is more likely to be able to organize itself than a larger one (because of the size of the stake and the homogeneity of interest) while anyone organized to achieve the "public good" will pay most of the costs and receive a smaller fraction of the benefits, special interest groups will more often than not act in support of efficiency-reducing measures. While each inefficiency-inducing measure (tariffs, farm price supports, tax loopholes, etc.) generates one-time costs, the accretion of special interest groups implies that growth rates will fall over time as more and more form and become more powerful.

Olson uses this model to examine, in a broad sweep, the economic growth of the world, with special emphasis on the period after the Second World War: German, Japanese, and French special interest groups were largely destroyed by the war, and their ineffectiveness and inability to get special interest actions helped speed these countries' growth. Britain, by contrast, emerged from the war with interest groups (unions, etc.) intact and thus had more growth-inhibiting policies. In the French case, Olson goes further to argue that "the deep divisions in French ideological life must owe something to the way one upheaval after another called into question the country's basic political and economic system. . . . Such smaller groups as trade associations and the alumni of the most prestigious schools have, as the theory predicts, more often been able to organize, but their effects on growth rates in the last few decades have often been limited" (p. 26).

Olson's special interest group formulation is stimulating. As he points out in the introductory footnote to his article, he took account of the substantive contributions of the conference in his book and not in the paper included in the Mueller volume. Nonetheless, for the reader interested in a quick (and less rich) introduction to Olson's ideas, Olson's paper in the Mueller volume is well worth reading.

The remainder of the book contains papers examining aspects of the Olson hypothesis on a cross-country basis, by an in-depth examination of the experience of selected countries, from a different perspective (as in the paper by Samuel Bowles and John Eatwell that challenges the "economic orthodoxy" of the Olson view without, at least to my mind, offering a constructive alternative), and by an effort by Mueller to relate Olson's work to public choice theory.

Although many of the empirical papers strongly support Olson's work, the most interesting and thought provoking are those raising qualifications or questions about the model. J. C. Asselain and C. Morrison, for example, examine two centuries of French growth and conclude that "special-interest groups manage to preserve their privileges in all except revolu [illegible]
tions," and that what is decisive as to their role is the ov [illegible]
economy: during crisis or recession, their strength [illegible]
whereas during periods of rapid economic growth, [illegible]
liberalization but are sufficiently compensated by the be [illegible]
they point out, such a simultaneous approach to the [illegible]
interest groups and economic growth would provide a [illegible]
why recession and prosperity tend to be self-reinforcin [illegible]

A second paper focusing on France, by Jean-François [illegible]
the role of the formation of the Common Market as a key factor [illegible]
the rules of the game within France and permitting rapid growth. He interprets the weakened position of special interest groups after the Second War to be the result of a political decision taken at a time when interest groups either could not resist or did not anticipate the consequences of the change. An essay by Franz Lehner examines the Swiss case as another exception to the Olson hypothesis: relatively high growth is accompanied by the existence of many small special interest groups. Lehner attempts to explain this phenomenon by examining the interaction among these groups: in Switzerland, the process of reaching agreement among them is so difficult that there is a great deal of inertia in the decision-making process and government interventions

are therefore fewer than in other countries. Lehner interprets his observations to imply that Olson's hypothesis needs to be amended to consider modes of interactions among interest groups.

Of the papers generally supportive of the Olson hypothesis, possibly the most imaginative is that of Peter Murrell. Hypothesizing that the age of an industry would have been independent of the "strength of inertial forces" in Germany after 1948, but a function of those forces in the United Kingdom, Murrell concludes that a reasonable test of the Olson hypothesis would be to examine whether comparative industry-specific growth rates (between Germany and the United Kingdom) differed systematically with the age of alternative industries. He posits that, with slower growth, all U.K. industries would grow more slowly than German industries by the same proportion (and thus be independent of age and the "strength of inertial forces") if Olson's theory were incorrect. Thus, each industry's relative growth rate is expressed relative to the average rate over all industries in each country. On the basis of this test, Murrell finds "significant" inertial forces in British industry. Although the results are suggestive, there are two important questions pertaining to his findings. First, it is only in 55–64 percent of the cases (with a large sample) that the relativities are as hypothesized (the proportion is higher when only 15 industries for 10–15 years are examined). Second, and probably more important, there are two reasons to believe that other phenomena associated with slower growth might account for Murrell's findings. Engel's law suggests that with approximately equal initial incomes, there would be more dispersion in industry growth rates in a more rapidly growing country (because demand for nonfood items would grow more rapidly). Also, to the extent that tradables production grows more rapidly than home-goods production with higher growth rates, there is another reason why the "inertial forces" attributed to interest groups might in fact be attributable to slower growth.

Space precludes further description of the contents. Suffice it to say that Olson's hypothesis as he sets it forth is persuasive, even if the supporting evidence he marshalls is somewhat impressionistic. The empirical essays in this volume are thought provoking: one comes away with the view that special interest groups are but one aspect of the political-economic process, albeit an important one, and that those concerned with the determinants of economic policy decisions still have a great deal to learn.

ANNE O. KRUEGER

World Bank
University of Minnesota

Reference

Olson, Mancur. *The Rise and Fall of Nations: Economic Growth, Stagflation, and Social Rigidities.* New Haven, Conn.: Yale Univ Press, 1982.

Capital and the Distribution of Labor Earnings. By MICHAEL SATTINGER. Amsterdam and New York: North-Holland Publishing Co., 1980. Pp. xvi + 281. $46 25.

In the second half of the 1970s, Michael Sattinger published a series of papers on various aspects of the distribution of labor earnings, beginning in 1975 with an article in *Econometrica* describing a model of the allocation of workers to jobs. In that paper he demonstrated how comparative advantage among workers affects the relationship between the distribution of abilities and earnings, a relationship that at the time had not been extensively studied. In the preface to the book under review, Sattinger reveals that he had actually developed the main ideas behind the 1975 paper some years earlier, while still a graduate student, but that his thesis research was on another topic because he "could not demonstrate the empirical relevance of the concepts" (p. vii). By 1978, however, in an article in the *Review of Economics and Statistics*, Sattinger was able to provide evidence of the existence of comparative advantage, using data on mechanical aptitude tests.

Capital and the Distribution of Labor Earnings is an outgrowth of this series of papers. His fundamental argument is that earnings should be viewed as factor prices, as a result of which, like all prices, they play an allocative role, so that "the distribution of earnings and the valuations placed on individual characteristics depend on the particular equilibrium of the economy" (p. 5). Building on his earlier published work, Sattinger's aim is to establish as his "major ideological point . . . that an individual's contribution to production cannot be related or attributed solely to his individual characteristics" (p. 6). The book begins, therefore (after a brief introduction), with a criticism of the human capital approach, followed by a review of literature related to Sattinger's work; the review is necessary, he claims, because the literature has "largely been ignored or slighted in other surveys" (p. 23). The main theoretical content of the book comes next: three chapters dealing with the inequality multiplier (the ratio of earnings inequality to skill inequality), the assignment of workers to jobs, and a theory of the earnings distribution based on differential rents. Four chapters of empirical work testing these theories, a further three chapters extending the analysis, and a summary of what has been achieved, and what remains to be done, complete the book.

Let me say at once that I am broadly sympathetic to Sattinger's stated view that many economists have tended to study the distribution of earnings in an incomplete, or inappropriate, fashion. Furthermore, the work Sattinger published between 1975 and 1979 represents an important contribution to the subject, which deserves the wider exposure his book should bring, notwithstanding the increased interest in comparative advantage, optimal assignment, and the distribution of earnings over the last few years. It must be said, however, that the demands of proselytizing occasionally cause Sattinger to become a trifle cavalier. Let me provide an example

Chapter 2 is, as I have mentioned, a critical analysis [illegible] theory as it relates to the distribution of earnings. The ch[illegible] comprising only nine pages, and presents the outlines of [illegible] one attributed to Mincer, the other to Becker and Chisw[illegible] that Mincer's model necessarily implies a horizontal su[illegible] number of workers at any given schooling level, by virtu[illegible] that present values of lifetime earnings are equalized in [illegible] and that the Becker-Chiswick model implies a horizontal de[illegible] it assumes that all types of labor can be measured in terms [illegible] efficiency units, so that the relationship between any two types of labor is fixed. Thus, Sattinger suggests, "the most interesting questions concerning the earnings distribution" are assumed away (p. 21).

There is undoubtedly a sense in which this is correct, but it is still reasonable to ask, for example, how far such a model can go toward explaining observed inequality in annual earnings. To be sure, it will not provide a complete explanation, and Sattinger's work provides a demonstration why, but I think the contribution of Mincer, Becker, and Chiswick is still an important one, and I find Sattinger's criticisms just a little ungracious. Blinder's (1976) critical review of human capital models, although admittedly not directly pertinent to the question of the distribution of earnings, is, by contrast, much more balanced, as is Rosen's (1973) paper.

The heart of Sattinger's book follows the brief discussion of human capital theory: the review of work that Sattinger feels has often been overlooked, notably Tinbergen's and Roy's, and the group of chapters on theoretical issues. Not all of this material is new (e.g., chap. 5 includes results from Sattinger's paper in the *Journal of Economic Theory* in 1977), but it is useful to see it collected, and connected, together in this way. Indeed, if any criticism is to be directed at these chapters, it is that the threads could have been drawn together more frequently than Sattinger chooses to do. Additionally, little attempt appears to have been made to relate earlier work on the subject to the results Sattinger derives, but these are relatively minor objections that some may view as mere matters of taste.

Chapter 4 develops a theory of the relationship between earnings inequality and inequality in productive abilities (the inequality multiplier). The model is an aggregate one and hence omits any consideration of the distribution of productive abilities. Instead, the existence of such a distribution is taken as given, production is specified as a function of capital intensity and average productive ability, and the effects of changes in these factors on earnings inequality are investigated. The approach is "essentially neoclassical" (p. 40): for the most part, earnings and marginal products are synonymous. Sattinger finds that the conditions under which the inequality multiplier does not vary with capital intensity and average productive ability are extremely restrictive, which provides the basis for his criticism of the assumption of Becker and Chiswick that labor can be measured in efficiency units. He also shows that the effects on earnings inequality of capital intensity and average productive ability are theoretically ambiguous, although if capital-skill complementarity is assumed, the inequality multiplier is positively related to capital intensity.

In the following two chapters, Sattinger turns his attention to the development of a disaggregated analysis. First, in chapter 5, he employs a sequence of increasingly sophisticated models to examine the equilibrium assignment of workers to tasks. Initially the maximization of a utility function, whose arguments are earnings and job satisfaction, guides individuals to jobs. Then, with individuals differing by ability and jobs by degree of difficulty, the existence of comparative advantage implies a technically most efficient assignment. Next, a scale-of-resources effect is introduced, first in isolation from, then together with, comparative advantage. The essential point of this effect is that those workers with greater productive ability are more valuable, relative to their less productive colleagues, if they are used in conjunction with a greater quantity of resources and are therefore assigned to jobs requiring greater resources. Throughout the chapter, particular emphasis is given to the conditions under which the more productive workers are used on more capital-intensive production processes; for example, Sattinger shows that, in the most sophisticated model, a sufficient, but certainly not necessary, condition for the result to occur is that the more productive worker has a comparative advantage in the more capital-intensive process.

Chapter 6 looks at the implications for the earnings distribution of the analysis in chapter 5. The model is "unabashedly short-run" (p. 117), in that the distributions of machine size and productive ability are assumed to be exogenous. These distributions, plus the production function and an assumption of an equilibrium assignment, then determine the wage and capital rent functions. In a series of illustrations using specific functional forms (e.g., a Cobb-Douglas production function and Pareto distributions of machine size and productive ability), the general result is established that earnings inequality is an increasing function of the inequality in the distribution of machine sizes.

The section of the book that follows the chapters on theoretical issues constitutes a fresh attempt to provide the empirical validation of Sattinger's ideas that earlier had proved so elusive. After a discussion of conceptual issues (chap. 7), successive chapters consider differences in U.S. earnings inequality over time, differences in inequality across states, and trends in wage differentials through time. Although one might question, while reading these chapters, the extent to which it is appropriate to refer to "tests of the hypotheses . . . developed in previous chapters" (p. 139) when the estimated equations are not typically *derived* on the basis of the earlier theory, the approach is undoubtedly novel, and Sattinger merits praise for his ingenuity in its formulation.

Chapter 8, for instance, comprises a set of time-series regressions with first the variance of logarithms of income[1] and then the shares of the top 1 percent and 5 percent of income recipients as the dependent variables; a similar analysis, this time cross-section, with the coefficient of variation of earnings as the dependent variable, is undertaken in chapter 9. The predictions of chapter 6 receive strong support; Sattinger's expectation of a positive impact on inequality of the capital to labor ratio is not, however, borne out by the empirical work. This expectation was based upon a strong belief in the existence of capital-skill complementarity, and Sattinger, in his concluding chapter, is refreshingly ready to acknowledge that the evidence from the empirical tests has caused him to revise his belief.

I do have a few criticisms of the empirical work of chapters 8–10. First, Sattinger claims that the results of chapter 8 "provide an answer to the question of why income inequality has failed to decline significantly
period, despite substantial increases in mean schooling
inequality in the distribution of schooling" (p. 152)
demand variables are counteracting the effect of th
Sattinger's analysis does, I feel, ignore a number of
issues, notably the influence of changes in the male to
force, and the effect of changes in the age structure
latter is, of course, the issue raised by Paglin (1975) in
the trend of income inequality. To be sure, Paglin's pape
its fair share of detractors, besides which Chiswick and Minc
original study (1972) that the effect of age changes is slight. Nevertheless, Sattinger's analysis is incomplete in failing to refer to the problem.

A second objection I have is that Sattinger repeatedly seems willing to interpret results a little generously when testing his hypotheses. Consider, for instance, the following: he refers to a coefficient that is "positive in all cases but is only significant in some regressions" as giving "mild support to the hypothesis of chapter 4" (p. 174), and later a coefficient that is "always positive but is not in general significant" is seen as lending "mild support" to another hypothesis from chapter 4. I should prefer the interpretation of the results to be more guarded than the use of the qualifying adjective "mild" seems to imply.

---

[1] Sattinger is adopting the data of, and reexamining the analysis of, Chiswick and Mincer (1972); like Chiswick and Mincer, he acknowledges that earnings data, if available, would be superior.

The remaining chapters of Sattinger's book extend his empirical investigations in a number of directions. Chapter 11 looks again at the impact over time of capital intensity on earnings inequality, but from a disaggregated viewpoint, the disaggregation being by industry, in the hope that this will overcome the problem of sectoral shifts' confounding the interpretation of the results. Chapter 12, too, is concerned with the distribution of earnings by industry and seeks to explain differences across industries in the proportion of workers in the upper and lower tails of the earnings distribution by factors such as capital intensity. Finally, chapter 13 investigates the implications of hierarchies for the distribution of earnings by a number of means; one of these is a set of regressions relating the Pareto coefficients from the upper tails of different industries' earnings distributions to various industry characteristics, among which is the ratio of production workers to total employment, as a measure of (the inverse of) the size of the hierarchy.

In his concluding remarks, Sattinger states that his original intentions in writing this book were to "change economists' minds about the nature and determinants of the distribution of earnings" (p. 267) and to provoke a "useful and scholarly debate" by an "aggressive presentation of differences" (p. 268). *There are, admittedly, occasions when I think I should have preferred* slightly less aggression; on balance, though, Sattinger is at least as stimulating as he is provocative, and I hope his book will be widely read.

Alan Harrison

McMaster University
University of Warwick

References

Blinder, Alan S. "On Dogmatism in Human Capital Theory." *J. Human Resources* 11 (Winter 1976): 8–22.

Chiswick, Barry R., and Mincer, Jacob. "Time-Series Changes in Personal Income Inequality in the United States from 1939, with Projections to 1985." *J P E.* 80, no. 3, pt. 2 (May/June 1972): S34–S66.

Paglin, Morton. "The Measurement and Trend of Inequality: A Basic Revision." *A E.R.* 65 (September 1975): 598–609.

Rosen, Sherwin. "Income Generating Functions and Capital Accumulation." Discussion Paper no. 73-15. Rochester, N.Y.: Univ. Rochester, June 1973.

In Memoriam: Harry Gordon Johnson
(1923–1977)

BACH

Harry Gordon Johnson was born on May 26,
tario. He was graduated from the University of
the age of 20 he began his first teaching post
University in Nova Scotia. His academic career
1944 by a year's duty with the Canadian infantry
when the Canadian Army was not able to send
immediately, he proceeded to Cambridge University
where he received a B.A. in 1946. He continued his education at Toronto (M.A. 1947) and Harvard (M.A 1948). In between he worked his way across the ocean in a cattle boat to visit friends at Cambridge. In 1949 he returned to Cambridge to teach (both at Jesus College and King's College) and received another M.A. in 1951. Disillusioned with the intellectual climate at Cambridge, he moved on to Manchester University in 1956. He was appointed Professor of Eco-

nomic Theory and taught there until 1959. During this time he received his Ph.D. from Harvard (1958).

In 1959 Johnson began his 18-year tenure at the University of Chicago, first as Professor of Economics and then as the Charles F. Grey Distinguished Service Professor of Economics (1974). He held simultaneous positions at the London School of Economics (1966–74) and at the Graduate School of International Studies in Geneva (1976–77). Throughout his career he held visiting positions at many universities around the world, including Karachi, Singapore, and Buenos Aires.

Johnson supplemented his academic duties by editing several academic journals. He served as assistant editor of the *Review of Economic Studies* from 1950 to 1959 and remained on its editorial board until 1965. He edited the *Manchester School of Economics and Social Studies* (1956–59), *Economica* (1969–70), and the *Journal of International Economics* (1969–76). He served as a series editor for Aldine Publishing Company (1964–77) and as Director of Research for the Trade Policy Research Centre (1968–77). He was the sole editor of the *Journal of Political Economy* from 1960 to 1966 and joint editor from 1969 until his death in 1977.

Johnson served on the boards of many professional associations, including the Association of University Teachers of Economics (chairman, 1968–71), American Economic Association (vice-president, 1976), Canadian Economic Association, National Institute for Economic and Social Research, Pakistan Institute of Development Studies, Royal Economic Society, Canadian Political Science Association (president, 1965–66), British Association (president, 1972–73), and Eastern Economic Association (president, 1976–77). He also held various public service posts, including the Royal Commission on Banking and Finance (Canada), the President's Special Group on the Great Society (United States), and the Council for Scientific Policy (United Kingdom). His numerous academic and public service functions did not adversely affect his own scholarly output.

Johnson received seven honorary degrees and fellowships in various societies, including the American Academy of Arts and Sciences, the British Academy, the Econometric Society, and the Royal Society of Canada. He also received the Prix Mondial Nessim Habif from the University of Geneva and the Bernard Harm Prize from the University of Kiel.

Gladstone's private secretary once said that "if a figure of 100 could represent the energy of an ordinary man, and 200 that of an exceptional man, Gladstone's energy would represent a figure of at least 1,000."* Harry Johnson was in the Gladstonian tradition.

---

* Philip Magnus, *Gladstone: A Biography* (London: Murray, 1954), p. xi.

Harry Johnson's Contributions to International Trade Theory

W. M. Corden

Australian National University

The object of this survey is to provide a thorough guide to Harry Johnson's work in the field of nonmonetary international economics—to "do a Harry Johnson" on that part of his work which includes many of his most influential contributions. He would certainly have wanted this particular survey to be thorough, as so many of his more original insights are to be found in this field. It is a field which he dominated over 2 decades, and not only through his writings.

The organization of the survey is by topics, with some attention to chronology and with inevitable overlap problems. Section I covers trade and growth, dominated by a trilogy of articles of the early 1950s, of which the third, published in 1955, is perhaps his single most significant contribution to trade theory. Section II covers static positive Heckscher-Ohlin-Samuelson (H-O-S) theory, where all his main contributions were the fruits of the 1950s, publis [illegible] *Manchester School.* Section III deals with no [illegible] theory—the theory of the optimal tariff, the cos [illegible] the theory of domestic distortions. His contribu [illegible] whole period of his working life up to the early 1 [illegible] defined to exclude the contents of Section IV, w [illegible] union theory and the theory of tariff bargaining. Sec [illegible] effective protection, in which his principal contributions [illegible]. from 1965 to 1971. Section VI is entitled "Wicksell Lecture and After" and deals with his writings from 1968 on the new theories of

---

I am indebted to comments on earlier drafts from Kym Anderson, Heinz Arndt, Stanley Engerman, Ronald Findlay, Ross Garnaut, Melvyn Krauss, John Martin, and Jonathan Pincus.

[*Journal of Political Economy*, 1984, vol. 92, no. 4]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9204-0004$01.50

trade, the role of technology, human capital, and knowledge diffusion in relation to trade, international corporations, and the infant industry argument. In fact, it is a bit of a potpourri, like his famous Wicksell Lecture, but deals with some of his most stimulating work, about which he was actively thinking in his last years. Section VII tidies up, dealing with the large volume of writings not covered to this point, notably several surveys, his work on (nontrade) two-sector theory—where he used the techniques of trade theory for income distribution and other issues—his writings on international trade negotiations, and on policies toward less developed countries. The last includes what is probably his most widely read book, *Economic Policies toward Less Developed Countries* (1967*a*). Finally, Section VIII discusses his attitude to theory and his contribution to the work of others in the field.

Before launching into details, we must say something about the essential nature of Harry's achievements and especially the matter of his "originality." His brilliance was as a synthesizer and a user of theory to probe "real world" policy issues. As a contributor to pure theory, he must be regarded as having had an important place in the ongoing stream: he said neither the first nor the last word in the many branches of trade theory in which he wrote, but he played a big role in the collective process of building up the relevant analytical framework. In total, his contribution to pure trade theory was immense, but his originality in this field was not of the order of Ohlin's, Samuelson's, or Meade's.

As we read his many papers, one characteristic stands out: he did his best to *appear* unoriginal. His style of writing and care in acknowledgments tended to give the impression that he had not had an original idea in his life but that everything was consolidation and building on bricks laid by others. He tended not to highlight what was new in his work but rather to stress continuity in the development of economic theory. He was not a deliberate maximizer of his reputation. He made no claims for original "theorems" and usually went out of his way to give credit to others. In the preface to *International Trade and Economic Growth* (1958*b*)—in which he collected some of those papers that have turned out to be his most important—he wrote about his work: "In addition, they are linked by a common purpose—the consolidation of the work of previous theorists and its extension to new problems—and by a common method—the application of mathematically-based logical analysis to theoretical problems thrown up by discussion of current topics among economists, generally those relating to economic policy. The motivation, explicit and implicit, has usually been to see what could be made of the tools available, in analysing the problem that suggested itself as an interesting one" (p. 9).

This impression of unoriginality was fostered not only by the many surveys he wrote but also by the substantial element of surveying in almost every article. It was fostered by his search for simplicity in theoretical argument. It is not always realized how difficult it is to be simple.

Because of this characteristic, I have made a special point here of noting the degree of originality of various papers and of digging the nuggets of originality from what appear to be expositional or survey articles. This seems to be a necessary role for a survey of this kind. But it should not obscure the point that the Great Synthesizer should not be judged solely by the criterion of theoretical originality.

I. Growth and Trade

It is appropriate to begin with the Johnson article that has had the biggest impact on trade theory, namely, "Economic Expansion and International Trade" (1955). This article has provided the basis for one or more chapters in every trade-theory textbook and has practically founded a subbranch of the subject. It took growth in two countries through factor accumulation and technical progress as given and spelled out the effects on the volume, pattern, and terms of trade in a two-good, two-country model. It introduced the concepts of protrade, antitrade, and ultra pro- (and ultra anti-) trade biases. The article itself is extremely rich and dense in one of the characteristic Johnson styles, so that it is hardly possible to summarize it here. It explores some cases and problems that have been lost in the subsequent literature. It was inspired by Hicks (1953) and also made use of [illegible] (1955), but there can be no doubt about its fundam[illegible] has provided the framework for analyzing a [illegible] lems, notably the effects of growth on the ter[illegible] oping relative to developed countries. It founded [illegible] which the most important subsequent contribu[illegible] Grubert (1959), which explored the effects of [illegible] progress in the model, a matter only hinted at in [illegible]

This major work was closely followed by two other [illegible] same field. "Effects of Changes in Comparative Costs as Influenced by Technical Change" (1961) analyzed the effects of technical change on trade in detail. This article deserves to be more widely known. It expounded the Findlay-Grubert analysis verbally with great clarity and in addition expanded this model: a nontraded good was introduced, linking the analysis with Hicks (1953), and perfectly elastic supplies of one of the two factors were allowed for. The question was then posed whether technical change is likely to inhibit or encourage international trade. Are government and private research likely to be import or export biased? The transmission of knowledge between countries was discussed. It is interesting to read this article now; it turns out to be the link between his early growth-and-trade work and the later themes of the Wicksell Lecture.

In "Economic Development and International Trade" (1959*a*) he "packaged" the whole growth-and-trade subject very neatly. This is the article on biases now read and referred to. In it he expounded the essence of his 1955 article with the aid of some neat diagrams, leaving behind the verbal density and excessive taxonomy of the earlier article; he incorporated the Rybczynski and the Findlay-Grubert analyses; and he gave the profession (as he would put it) a simple article embodying all the main ideas in a way that a reasonably attentive undergraduate could understand.

The classic 1955 article was the third of a trilogy of articles on growth and trade published in the 3 years 1953–55. The first, "Equilibrium Growth in an International Economy" (1953), extended the Harrod-Domar analysis to the open economy. But Harry recognized subsequently that "the crux of the Harrod-Domar analysis, the knife-edge requirements of self-justifying growth, has, however, ceased to interest economic theorists, in consequence of the recognition that the savings ratio and the capital-output ratio are not parameters but variables amenable to economic policy or subject to equilibrating forces" (1958*b*, p. 10).

The second article, "Increasing Productivity, Income-Price Trends and the Trade Balance" (1954*a*), was very influential in Britain at the time. It had been commonly argued that faster productivity growth in the United States than in European countries would tend to improve the U.S. balance of payments (the "dollar problem"). He subjected this to rigorous, indeed elegant, analysis, introducing income as well as price effects—showing that income effects were likely to worsen the balance of payments of the faster-growing country. His monetary assumptions in this article were rather simple—he made various explicit money-income assumptions, with the required monetary policy implicit. While they were not unreasonable for the purpose of the exercise, I doubt that he would have approved of this approach in his later "monetary theory of the balance of payments" days. Furthermore, a complete analysis would really have had to introduce the complications of the 1955 article, allowance being made for more than one good produced in each country.

Something must be said about the quality of the exposition in these two articles. In both cases Harry expounded difficult arguments superbly. He used simple mathematics, avoided excessively complicated geometry, and clearly explained all steps verbally. Young theorists who wish to be understood by their readers, encouraged by editors who wish their journals to be read, would benefit from following the expository techniques of these articles, especially the "Increasing Productivity" article.

The 1953 article contained a rigorous statement, with appropriate formulae, of the concept now known as "immiserizing growth"—the case where factor accumulation or technical progress can lower a country's aggregate real income because of a sufficiently adverse effect of growth on its terms of trade. Awareness of the possibility goes back to Mill and Edgeworth, but Harry was the first to revive the case and explore it in rigorous modern terms. He did not give it a name or a diagram, however, and he did not highlight it. In this case he left an adequate exposition to his pupil, Jagdish N. Bhagwati, whose article is now rightly the standard reference on the subject (Bhagwati 1958). Furthermore, the Edgeworth-Nicholson controversy on the subject was sorted out neatly in Bhagwati and Johnson (1960).

Harry later came back to immiserizing growth in two short notes. In "The Possibility of Income Losses from Increased Efficiency or Factor Accumulation in the Presence of Tariffs" (1967*d*), he showed in the usual two-sector general equilibrium model that technical progress or capital accumulation could bring about a fall in real income even when the terms of trade do not change—that is, in a small country model. This becomes a possibility when technical progress is concentrated in the protected industry, or the importable is relatively capital intensive and capital accumulates, the cost of protection rising so much as to outweigh the income gains obtainable in free trade. This possibility was actually first uncovered by Dales (1966), who expounded it in partial equilibrium terms and whom Harry acknowledged. In "A Note on Distortions and the Rate of Growth
Economy" (1970*c*), the problem is to consider the
the rate of growth when there is given capital acc
cal progress, with the import-competing industry
the technical progress biased toward the import-
His "The Possibility of Income Losses" (1967*d*)
would lower the rate of growth in this case becau
effect of the growing cost of protection. On the other
might expect that the more production is shifted toward the industries intensive in the faster-growing factor, the greater the weight of that factor's growth rate in the total growth rate and hence the greater the rate of growth. Harry found that the growth result could go either way, though he did not succeed in obtaining any simple results. This issue deserves further study.

In 1971 Harry published two further substantial, though overlapping, articles on trade and growth, one in the Kindleberger festschrift (1971*c*) and the other in the *Journal of International Economics* (1971*d*). His aim was to extend the static H-O-S model to include endogenous capital accumulation with a constant savings ratio and constant population growth. He built on the work of a group of "more mathematically-inclined theorists," notably Oniki and Uzawa (1965), though he certainly brought out effects that are not clear in their articles. On the other hand, he limited himself to the small country case in order to make the problem tractable with his verbal and geometric methods. His explicit aim was to make this broad type of analysis more widely accessible—"to alleviate this undesirable situation of mathematicians' monopoly."

As usual with Harry's articles, one has to dig out the elements of originality (apart from the basic geometric exposition): these were his analysis of the effects of a tariff on long-run growth equilibrium (in the small country case) and the incorporation of increasing returns to scale into the comparative static analysis of the effects of growth on trade, an extension of his 1955 article. One can have reservations about the exposition in these two articles. A less busy writer would have gone on to produce the equivalent of "Economic Development and International Trade," in which the diagrams become simple and the main points emerge clearly without tiresome taxonomy. One also wonders why he wrote these papers, since he did not believe in the value of this (Oniki and Uzawa) body of mathematical trade theory. In the Wicksell Lecture (1968*a*) he wrote about it: "The kinds of problems suggested for theoretical analysis by this procedure, however, are too remote from the real world to be very interesting" (p. 7) I suspect he could not resist the temptation to clarify an intricate body of theory, if only to show that he had mastered it.

II. Positive Static Trade Theory

During his years at Manchester University from 1956 to 1959, Harry produced three articles that represent his principal contributions to mainstream H-O-S theory (other than the growth-and-trade work discussed above) and that have become standard references, as innumerable citations to them testify.

He consolidated the two-country, two-factor Heckscher-Ohlin model in "Factor Endowments, International Trade and Factor Prices" (1957), bringing out clearly the possibility of factor reversals. It was inspired by Harrod (1958) and built around a diagram originating in Samuelson (1949). This is one of his most widely read and cited articles in pure trade theory. In his own words, its function was "to review the existing literature with the double object of verifying the accuracy of accepted conclusions and synthesizing the methods and results of previous writers into simpler, more readily usable analysis" (preface, *International Trade and Economic Growth* [1958*b*, p. 9]). The possibility of factor reversals had been uncovered by Pearce (1951–52) and further explored by Jones (1956). In a sense, this paper was no more than an improvement on Harrod (1958), and he acknowledged this. What he did was to package the whole subject so neatly and tidily, with such economy of words and utter clarity, that his article, rather than the various earlier papers, is still read and has provided the basis for textbook writers.

"International Trade, Income Distribution, and the Offer Curve" (1959*b*) and "Income Distribution, the Offer Curve, and the Effects of Tariffs" (1960*c*) introduced into the standard H-O-S model the effects of income distribution changes, through resulting changes in the aggregate pattern of demand, on the offer curve. Although H-O-S theory had brought out the effects of trade and tariffs on factor income distribution, it had not allowed for the further repercussions of changes in income distribution on demand owing to differences in the spending patterns of factor owners. This introduced new possibilities of instability and multiple equilibria. These two articles—which were later republished as one in *Aspects of the Theory of Tariffs* (1971*a*)—filled an important gap in H-O-S theory. They were tightly argued, a feast of verbal mathematics and intricate theory. The second article incorporated the Lerner (1936), Metzler (1949), and Bhagwati (1959) cases concerning unexpected effects of tariffs on the terms of trade, on domestic prices, and on factor incomes; hence it was also a work of consolidation.

In a joint article with Bhagwati, "A Generalized Theory of the Effects of Tariffs on the Terms of Trade" (1961), he followed up this theme. The two authors showed that tariffs may worsen the terms of trade by reducing the demand for imports through a variety of unexpected spending effects, the government and pri
separated. The various cases are carefully wo
pounded with great clarity, though most are, p
relevance to any practical issue. But this articl
1959 and 1960 articles, tidied up this whole body
income distribution is affected by changes in rel
turn, through effects on demand, influences the

Harry believed in the two-sector model because h
simplicity in theory and because he was, above all, a teacher. He learned to manipulate it for numerous purposes. The two articles on income distribution and the offer curve (1959*b*, 1960*c*) provided the foundation for much of his later two-sector work, discussed further in Section VII below.

Harry's only other contribution to H-O-S theory was in two short notes on factor-price equalization when commodities outnumber factors (1967*c*, 1970*d*). The contribution of these notes was not just to reaffirm the proposition of Samuelson (1949) that the introduction of extra commodities relative to factors strengthens the likelihood of *factor-price equalization—and certainly that factor prices cannot actually diverge because of this—but to bring out the role of demand* conditions for this problem: "demand conditions may make factor prices under trade converge less in the three-good than in the two-good case. . . ."

III. Normative Theory of Protection

In this field Harry wrote six substantial theoretical articles, including a trilogy on the optimum tariff and two of his most important articles—on the cost of protection and on domestic distortions. In addition, there is a pioneering empirical piece estimating ex ante the gains to the United Kingdom from freer trade with Europe.

His first contribution to the theory of protection was "Optimum Welfare and Maximum Revenue Tariffs" (1951–52), where he showed that the maximum revenue tariff must always be higher than the optimum tariff and also produced a neat consolidation of various optimum tariff formulae. The latter part was reprinted in *International Trade and Economic Growth* as "Alternative Optimum Tariff Formulae" and is a standard reference. The theory of the optimum tariff from a national point of view (taking into account the terms of trade effects of a tariff) was already fully developed by the time Harry moved into the field. His article helped to tidy it up.

"Optimum Tariffs and Retaliation" (1954*b*) was more significant and ambitious. He set out to "reassert the proposition that a country *may* gain by imposing a tariff even if other countries retaliate; and to determine the conditions under which it *will* gain in a special group of cases." This was done, typically, with great thoroughness. His starting point was the classic article by Scitovsky (1942) where the framework of the approach had been set out, and Harry was really concerned to expose an error in this article (that, once retaliation occurs, *all* parties are bound to lose). He clearly succeeded in doing so, but the real question must be about the way in which the retaliation process was conceived. Once country A imposes an optimum tariff (assuming no tariff by B), B then retaliates with an optimum tariff (assuming no change in A's tariff). Country B's retaliation consists of imposing an optimum tariff, not a bargaining tariff; and neither country takes into account the possibility that its own tariff will lead the other country to change *its* tariff. In a footnote Harry showed he was aware of this difficulty, but while a case can perhaps be made for this type of model in a multicountry model and when certain types of lag are allowed for, one must doubt whether so elaborate an analysis was worthwhile.

Harry's third contribution to optimum tariff theory came much later, with "The Gain from Exploiting Monopoly or Monopsony

Power in International Trade" (1968*b*). He showed that the national gain from an optimum tariff (without retaliation) would be higher the greater the domestic elasticities of substitution on the supply-and-demand side and the higher the ratio of imports to national income. On plausible figures he showed the likely gain to be small.

"The Cost of Protection and the Scientific Tariff" (1960*a*) was among Harry's most important articles. It marked a break with the general lines of the "orthodox" trade theory and was the first of a number of contributions from him that have helped to make the theory of trade policy much more relevant. The first part of the article, influenced by Corden (1957), developed a rigorous cost-of-protection theory in a multigood general equilibrium framework. His improvement on my work was to go beyond two goods and to spell out clearly the general equilibrium implications. The second part, on the "scientific tariff," was influenced by Young (1957). He worked out the minimum-cost tariff structures required to achieve various non-economic objectives, his aim being to formalize the logic of actual tariff policies. A development of this article was "The Costs of Protection and Self-Sufficiency" (1965*a*) in which he attached magnitudes to the consumption cost and the production cost of protection (compared with free trade) on the basis of plausible social utility functions and transformation curves. He granted that "the model . . . is far too simplified to permit the drawing of any very firm conclusions about reality, . . ." but concluded, among other things, "that the costs of protectionist policies are unlikely to be large enough, relative to potential maximum national income, to account for much of the existing major differences between the real per capita nat
various countries" (p. 371).

"Optimal Trade Intervention in the Presence
tion" (1965*c*) has proved to be one of Harry's
more so, possibly, than his 1960 "Cost of Prot
greatly influenced my own work, I can hardly be
He showed that in the presence of domestic distortions
never first best, and that if trade taxes are used to eliminate the original distortions a net loss may ensue. He worked systematically and brilliantly through the principal arguments for protection, notably arguments resting on factor immobility or factor-price rigidity and the infant industry argument. It is true that the basic idea could be found earlier, in Meade (1955), Hagen (1958), and Bhagwati and Ramaswami (1963). Harry acknowledged as his principal inspiration the last article ("to these two authors belongs the credit for reducing a mass of ad hoc arguments concerning tariffs to a simple application of second-best welfare theory"), but any comparison with this paper, the article by Hagen, and the book by Meade must show how much of an advance Harry's article made on its predecessors. Apart from the

beautifully expounded details, he provided the correct focus. Naturally, the subject has advanced since then, partly through rediscovery of themes in Meade (1955) and clarification of various implicit assumptions of "domestic distortions" theory. Harry himself later looked again at the infant industry argument.

Finally, mention must be made of his one empirical article in this field, "The Gains from Freer Trade with Europe: An Estimate" (1958*a*). Here he applied simple Marshallian (partial equilibrium cost of protection) analysis to a current problem, using ex ante data produced by Britain's Economist Intelligence Unit. The approach was bold and quite simple in a way that has subsequently become fashionable, but the fact is that it was one of the first calculations of this type, an advance on the "cash cost of protection" calculations previously made in Australia and Canada. Hence it is an important, much-cited article.

IV. Customs Union and Tariff-bargaining Theory

Harry's principal contribution to orthodox (Vinerian) customs union theory was an apparently expository article, "The Economic Theory of Customs Union" (1960*b*), which has been widely read and cited mainly because of the simplicity of its exposition compared with other papers available at the time. But it contained two novel features. First, he briefly worked out a case with economies of scale (foreshadowed, but not fully developed, in Viner [1950]). Second, he redefined trade creation to consist of a production *and a consumption* effect—instead of adhering to the Meade-Lipsey usage, in which trade creation referred only to a shift in the pattern of production leading to extra trade, while consumption shifts yielding extra trade were regarded as additional. The Meade-Lipsey approach meant that there could be a welfare gain from a trade-diverting customs union, the gain resulting from an offsetting consumption effect. It seemed much more sensible to Harry to combine all the effects that create trade—and thus yield welfare gains in the simple models used—and he set these off against the welfare-reducing effects of trade diversion. It is surprising that such a simple insight has not been more widely accepted. Harry came back to this point later, in "Trade-diverting Customs Unions: A Comment" (1974*b*), when he felt that unnecessary complications were being created through an inelegant definition based (probably) on a misinterpretation of Viner. In "A Note on Welfare-increasing Trade Diversion" (1975, p. 117) he argued that "this possibility is a needless analytical complication, since any such gain [i.e., gain from a union that was 'trade diverting'] must involve an increase in the volume of trade." Nevertheless, in the latter article he went on to work out the empirical possibility of such a case (on the basis of plausible utility and production function figures and an initial 25 percent tariff) and concluded that it was "probably an empty theoretical box."

"An Economic Theory of Protectionism, Tariff Bargaining, and the Formation of Customs Unions" (1965*b*) was one of Harry's most ambitious articles, and he regarded it as one of his most significant. His aim was to explain the logic of tariff bargaining and of the formation of customs unions, using the idea that "industrial production . . . appears as a collective consumption good yielding a flow of satisfaction to the electorate independent of the satisfaction they derive directly from the consumption of industrial products" (p. 258). Industrial production is a public good, and tariffs are used to foster industrial production because of a convention against export subsidization. On this basis he built an extremely elaborate theory designed to explain tariff bargaining, including tariff structures—notably, the potential benefits from discriminatory reciprocal tariff reductions. With regard to the formation of customs unions among developing countries, the approach seems useful and was developed at the same time more thoroughly in Cooper and Massell (1965). But with regard to the part of the article that was more original to Harry—the part that was concerned with the explanation of tariff bargaining among industrial countries—the hypothesis of "industrial production as a collective consumption good" has always seemed implausible to me.

The logic of tariff negotiations and tariff bargaining interested Harry over many years, and in various articles he explored different approaches. In "Optimum Tariffs and Retaliation" (1951*b*) he built on the optimum tariff approach. In "The Cost of Pro
Scientific Tariff" (1960*a*) he briefly analyzed t
"bargaining tariff" "aimed at inflicting econom
other country or countries in order to obtain
concessions" (p. 344). Then there was the app
under discussion, where industrial production
sumption good, and a country's aim in bargain
exports of industrial products for extra imports
back to the logic of reciprocity in "Trade Negotiations and the
International Monetary System" (1976*b*), where he favored an explanation of bargaining policies in terms of a balancing of domestic effects within each country—damaging effects of extra imports on particular import-competing sectors being set against expected gains for exporters and consumers. "Further, what is influential politically is not the change in the volume of production, but the number of people and managers sufficiently affected either adversely or favourably by that change to motivate them to try to influence government policy" (p. 21). This is by far the most plausible framework and gives

an appropriate weight to the reluctance of governments to take explicit action that will damage particular sections of the community. Clearly, had Harry lived he would have developed this line of thought further and might neatly have consolidated the various possible approaches to explaining negotiating processes.

Another article that might be noted at this point is "The Implications of Free or Freer Trade for the Harmonization of Other Policies" (1968*c*). This dealt exhaustively with an issue then and now relevant in the European Economic Community (and also in Canada, with regard to a possible free-trade area with the United States). It must now be the standard reference on the subject. His conclusion was that "a free trade area does not in and of itself require extensive harmonization of other policies. . . ." ". . . such harmonization is more a matter of choosing to augment the benefits of free trade than of being required to harmonize as a result of free trade."

V. Effective Protection

Harry will always be known as a pioneer, as well as popularizer, of effective protection theory. He wrote five articles in this field. His pioneering contribution was based on a lecture delivered in Geneva, "The Theory of Tariff Structure, with Special Reference to World Trade and Development" (1965*d*), in which he explored many aspects, characteristically correcting an error or imprecision in Barber (1955) concerned with the vertical escalation of tariff rates. This Geneva lecture was the first really theoretical article on the subject, closely followed by my own (Corden 1966). Apart from a late article, his writings on this subject were always rather piecemeal, and in this case he left the consolidation to me.

"A Model of Protection and the Exchange Rate" (1966*b*) brought out rigorously a paradoxical general equilibrium implication of an effective protection model—"the possibility that tariff structures may bring about a situation in which appreciation rather than depreciation would be necessary to preserve equilibrium under trade liberalization." "Nominal Tariffs, Indirect Taxes and Effective Rates of Protection: The Common Market Countries 1959" (1967), written with Herbert G. Grubel, was an important contribution, extending effective protection theory and making calculations to include excise taxes. The actual method used was subject to some valid criticisms from Gamir (1971), but this does not detract from the general significance of this article. "The Theory of Effective Protection and Preferences" (1969*b*) was, to some extent, an act of popularization and drew extensively and explicitly on my work; its particular contribution was to integrate effective protection theory with the theory of preferences, an interest that came out of his work on economic policies toward less developed countries. Finally, "Effective Protection and General Equilibrium Theory" (1971*b*) is a broad, very condensed survey of all the main general equilibrium issues raised by effective protection and is much influenced by my own work. Harry's own contribution—introduced characteristically without warning—is concerned with the pure trade theory case where there are more commodities than factors. He has difficulty reconciling this with the usual effective protection approach. He struggles with it, looking for simplicity with the aid of his favorite "Lerner-Pearce" diagram, and almost gives up, concluding realistically "that the prevailing mixture of general and partial equilibrium analysis is the most reasonable approach for practical economists interested in commercial policy questions, in spite of its rather shaky foundations in pure international trade theory."

Harry always acknowledged the Harvard doctoral thesis of William Travis as the source of his inspiration in this field, though he must have been influenced by Barber's (1955) article—"Canadian Tariff Policy"!—which developed the effective protection concept. One of the earliest empirical works—calculating "effective protection for labor" for the United States—came from his Chicago student Giorgio Basevi (1966). Harry's contribution was to show early awareness of the importance of the concept, to spread the simple idea and this awareness in the way he knew so well, with an emphasis on the idea's relevance for international trade negotiations, and to sort out some particular theoretical points.

VI. Wicksell Lecture and After

In his Wicksell Lecture, "Comparative Cost and C
Theory for a Developing World Economy" (196
break away from orthodox international trade th
who had been accustomed to tying up every em
assumption, this lecture was a departure He wan
neofactor proportions and the neotechnology
take into account international movements of ski
drain), the international corporation, and the much dis
time) technological gap between the United States and Western Europe. His purpose was "to attempt to synthesize the recent developments in the theory of international trade . . . in a broader approach to a dynamic theory of comparative cost and of commercial policy—in a developing world" (p. 8).

The result was a stimulating *tour d'horizon* of trade theories—an attempt to spell out the logic of (and develop further) the Posner-Hufbauer-Vernon theories of technological gap trade and of the product cycle, and much else. Apart from a superb survey and systematization of the new theories of trade, he developed his basic theme—the application of a generalized approach to capital accumulation whereby not only the accumulation of physical and human capital but also the growth of technology and knowledge in all forms are parts of a capital accumulation process. This takes place unevenly among countries—for reasons he discussed in detail—and then leads to a number of different processes of knowledge diffusion or transfer of production from the country of innovation to foreign locations. A subsidiary theme—which formed the basis for two important articles later—was that the extent of investment in knowledge and its diffusion would be socially suboptimal because of its well-known nonappropriability and public good characteristics.

There is no doubt that Harry valued this lecture highly, as evidenced by the references he made to its main arguments on many occasions later. In terms of its breadth, the issues it touches on or opens up, the sense of relevance that pervades it, and the stimulus it should provide for theorists and empirical workers in the future, it is a major work.

Admittedly, it is not rigorous. Too much is left loose. Harry's aim was to produce a grand, dynamic Heckscher-Ohlin model with the original factors—labor and nature—combined with capital in all its forms. Yet it seems unsatisfactory to treat all forms of capital accumulation—including investment in R & D—as in some sense similar. It is important for the pattern of trade what sort of capital is being accumulated. It has been shown that human and physical capital must be kept separate if the U.S. pattern of trade is to be explained (Branson 1977). Of course, Harry was aware of this, as of almost anything in this field that other scholars might think of. But this lecture cries out for more theoretical work, and the fact that so much insight and potential for the development of international economics had to be left loose is itself one component of the tragedy of Harry's early death. The reader is also distracted by a whole range of side issues relevant but not central to the main theme. One feels that Harry's mind was filled with thoughts in this field and, breathlessly, he had to spill them all out. Nevertheless, the Wicksell Lecture is a classic in international economics—an imaginative paper that showed Harry at his best as a man of broad vision, able to extract order out of somewhat chaotic theoretical developments and piecemeal empirical work. It is a paper that should inspire all laborers in this vineyard, especially graduate students in danger of getting the impression that trade theory consists mainly of arid model building.

Three substantial articles form part of Harry's "Wicksell period," namely, "Some Economic Aspects of Brain Drain" (1967*e*), "The

Efficiency and Welfare Implications of the International Corporation" (1970*a*), and "A New View of the Infant Industry Argument" (1970*b*). In addition, he really let himself go in one of his last papers, a contribution to the Nobel Symposium on the International Allocation of Economic Activity, "Technology, Technical Progress and the International Allocation of Economic Activity" (1977*b*).

The brain drain article is a brilliant, comprehensive, compact survey of all the issues: Canadian, British, and developing country aspects are covered; taxation, income controls, and externalities are allowed for; and unsound arguments are demolished. No stone is left unturned. This must be a standard reference in the economics of migration. His main conclusion is that the brain drain need not be harmful (though it *might* be). The paper is notable for a personal testament: "I adopt a cosmopolitan liberal position, and regard nationalism as one of the less pleasant mental vices in which mankind indulges itself, or as one of the characteristics of childish immaturity out of which I hope the people of the world will ultimately grow" (1967*e*, p. 379).

The much-cited "International Corporations" paper is built around the idea that the primary function of the multinational enterprise is the generation and international transfer of productive knowledge, and that "it does not necessarily follow that recompense through monopoly profits on the use of the knowledge is the ideal arrangement for the host country" (pp. 10–11). He was seeking a rationale for the various dissatisfactions of host countries with foreign multinationals. As a Canadian, he had long thought about these dissatisfactions. The paper also contains a detailed analysis of various argum [illegible] stricting inward foreign investment, though it p [illegible] various popular considerations that he would h [illegible] economic."

"A New View of the Infant Industry Argume [illegible] the same theme: ". . . the essential nature o [illegible] argument is now clear. What is involved is an in [illegible] of acquisition of knowledge which is socially pro [illegible] unprofitable because the private investor cannot appropriate [illegible] whole of the social return from his investment" (p. 60). When knowledge is a private good, there is no infant industry argument. The protection issue arises when "the relevant learning process has the inherent character of a public good" (p. 67). Traditionally, the infant industry argument has been defined as an argument for temporary intervention to protect a new venture, irrespective of the source of the conditions which justify intervention, as in his "Domestic Distortions" paper. But here he sought to narrow the definition—the essential feature of infancy being ignorance, and the argument being confined to grounds which relate to the creation and diffusion of knowledge. Harry explored here only one aspect of the traditional infant industry argument (as was pointed out in comments on this paper by Snape [1970]). But, within these limits, it is an important paper.

Finally, the Nobel Symposium paper really defies summary. He sketched out imaginatively a stylized historical process designed to explain the location of economic activity and its changes over time. Labor, capital, and technology are all mobile in varying degrees. He came back to the technological gap issue and why it is sometimes uneconomic to apply the "best" technology, and he stressed again the Wicksell point that "technology is a form of capital, and its availability alterable by investment." This must be regarded as a most stimulating, but incomplete, paper, and I hope it will be taken up and developed further by international trade theorists.

VII. Other Writings

We have now surveyed Harry Johnson's major contributions to nonmonetary international trade theory, yet an incredible number of publications remain to be noted—some of extremely high quality, including several surveys and his most widely read book.

Surveys

Harry was, of course, the surveyor general of the subject. In almost every article he surveyed the relevant literature before launching into his own contributions. Sections of the Wicksell Lecture stand out as a survey of the new theories of trade. Apart from this, his three most notable surveys are "International Trade Theory and Monopolistic Competition Theory" (1967*b*), the survey of international trade theory in the *Encyclopedia of the Social Sciences* (1968*d*), and "The Standard Theory of Tariffs" (1969*a*).

The "Monopolistic Competition" survey was a conscientious attempt to dig up everything that had ever been written on monopolistic competition and international trade—and he uncovered a surprising number of articles. But his main conclusion was that the theory of monopolistic competition had made virtually no impact on the theory of international trade, mainly because it was partial equilibrium and "appears to have no macroeconomic implications essentially different from those inherent in the assumption of pure competition habitually employed in international trade theory" (p. 204). However, he repeatedly came back to this issue, notably in the Wicksell Lecture and related writings, and stressed the relevance of monopolistic competition and product differentiation to the Wicksell issues.

The encyclopedia article is a superb piece of writing, showing Harry's ability to verbalize complex theoretical ideas, his historical perspective, and his sense of proportion. In spite of subsequent developments in the field—some of which he foreshadowed in brief remarks—it still repays reading and ought to have been reprinted in one of his collections.

"The Standard Theory of Tariffs" is a concise, very carefully written survey of the theory of tariffs as it was before the new ideas of the 1960s—effective protection, domestic distortions, and so on. It contains some new geometric devices. Harry said in the preface to *Aspects of the Theory of Tariffs* that the standard theory "struck me on reconsideration for this book as less scientifically valuable than the prior work on which it had been founded," but it is a pity he did not reprint it. Another well-known survey is "Tariffs and Economic Development" (1964*b*), which was updated somewhat in another survey, "Commercial Policy and Industrialization" (1972*a*), a lecture delivered in Panama and particularly directed to small countries. Finally, "Direct Foreign Investment: Survey of the Issues" (1972*b*) is a tight and comprehensive survey of the foreign investment issues from both a host and a home country point of view, much influenced by his lifelong involvement in the Canadian debate and also by Caves (1971). An original section deals with the effects of capital inflow on real wages, showing that these *might* fall. There is no better concise survey of the subject.

Two-Sector Model

International trade theorists pioneered the de
two-sector general equilibrium models, and it b
some time that their techniques could be app
trade theory—as evidenced by Meade (1965) an
tions to two-sector growth models. Harry reali
this, with one short book, *The Two-Sector Model*
(1971*e*), his joint text with Melvyn Krauss, *Genera*
sis: A Microeconomic Text (1974), and, in addition, a numbe
out articles on "Factor Market Distortions and the Shape of the Transformation Curve" (1966*a*), "On the Shape and Location of the Production Possibility Curve" (1973) with Krauss and Skouras, "A Note on Factor-Income Distribution and the Factor-Owner Production Block" (1976*a*), "A Note on the Geometry of the Two-by-Two General Equilibrium Model" (1977*a*), and "Factor Intensities and the Shape of the Production Possibility Curve" (written with Lizondo and Yeh in 1977 and published in 1981). The 1966 article generated a considerable, if rather esoteric, literature. It really belongs to the *positive* theory of domestic distortions and hence has been rather hard

to place in this survey. A common feature of much of this work involved the use of the Lerner-Pearce diagram and the development of some complicated geometry around it. Harry believed in the value of geometry as an expository device to spread economic literacy. He put this case strongly in his *The Two-Sector Model of General Equilibrium.* "Second, and more important, geometry lends itself far more readily than mathematical analysis to the communication of economic analysis to masses of students (though this advantage may be eroding as students in general become more numerate)."

A central theme of Harry's "two-sector" writings was a concern with income-distribution effects. His work in the field of income distribution falls outside the scope of this survey, but it was a valuable byproduct of his facility with general equilibrium two-sector models derived from trade theory. His most influential article here is "The Effects of Unionization on the Distribution of Income: A General Equilibrium Approach" (1970), written with Peter Mieszkowski. As mentioned earlier, much of this work might be regarded as a continuation of his two classic articles on income distribution and the offer curve (1959*b*, 1960*c*).

International Economic Relations

Harry's most widely read book is probably *Economic Policies toward Less Developed Countries*, written at the Brookings Institution in 3 months. It is a tour de force. Magpie-like, he picked up bits and pieces of the latest relevant research and adapted and fitted them all into a coherent framework. Utterly rigorous, clear, and balanced, the book covers numerous aspects, including the political, of the subject. It was his most successful venture in "spreading economic literacy." He did not avoid complex arguments and exhaustive analysis, but, aside from some footnotes, the exposition is all verbal. Spread through this book are many insights, but perhaps the principal contribution is the analysis of the UN Conference on Trade and Development (UNCTAD) scheme for temporary trade preferences for manufactures, which he showed to be a version of an infant industry argument for protection and also related to the theory of effective protection. The book is sympathetic to the UNCTAD ideas and motivations. In later years, Harry lost patience with the rather woolly and extravagant proposals described as "The New International Economic Order" and wrote several critical, even fierce, papers on the subject, reflecting particularly on the class interests of international bureaucrats.

In addition, Harry wrote numerous papers, overlapping with each other in content, on international trade negotiations, "alternative trade strategies for the seventies," the trend to protectionism, and so on. For a short time he became an advocate (rather than an analyst) of a particular scheme, namely, the proposals for a multilateral free-trade area. *The World Economy at the Crossroads* (1965*e*) brought together much of his thinking in this and related fields at an early stage. It went into three impressions. Always in these writings he maintained high standards of rigor and clarity. His last paper in this area, "Trade Negotiations and the New International Monetary System" (1976*b*), reflected on the effects of flexible exchange rates on trade negotiations.

VIII. Some General Reflections

What did Harry believe about the role of theory and about the state of international trade theory in particular? Did his own work conform to his stated beliefs? These are interesting issues in themselves, but they have been highlighted recently by the view expressed in Lipsey's (1978) survey of Harry's contributions to trade theory, namely, that Harry believed in "the emptiness of qualitative economics."

There is really not much difficulty in determining what Harry believed—he expressed his views clearly and forcefully in print on many occasions. The best statement can be found in "The State of Theory in Relation to the Empirical Analysis" (1970*e*). The relevant passages are too long to quote fully here, but he distinguished between (1) the role of theory in the context of empirical research and (2) the broad role of theory "as a systematic approach to the understanding of economic phenomena and as the organization of disciplined thinking about these phenomena and about policies relating to [illegible] first role its function is "to cast up empirically t[illegible] explanatory hypotheses, and the value of a theo[illegible] its explanatory power in comparison with its [illegible] role, the purpose is "to abstract from the comple[illegible] a simplified model of the key relationships betw[illegible] independent variables, and to explore the pos[illegible] implications of changes in the 'givens' of this hypoth[illegible] pointed out that "the theory of international trade has . . . [illegible] primarily theory in the second sense" (p. 10).

Harry's view of the role of models was well stated in his last year "... the purpose of models, and of teachers in general, is to instruct students in the simplest elements of a problem. Once they have that basic understanding of interconnections, the next stage is to look at the real world and try to see the simple interconnections operating" (Ohlin, Hesselborn, and Wijkman 1977, p. 99). At the same time, he was well aware of the need to adjust models in the light of new empirical evidence. This comes through strongly in "The State of Theory in Relation to the Empirical Analysis" and, above all, in the Wicksell Lecture.

In his later years, Harry became impatient with the "pursuit of arcane analytical problems." He felt that the trend to mathematical theorizing was discrediting trade theory among the general body of economists and preventing its wider application to economic policy. One of his earliest references to this view was in his review of Staffan B. Linder's *Essay on Trade and Transformation* (1964*a*): "This shift of emphasis is overdue, as is a recasting of the mathematically overelaborated present framework of trade theory" (p. 86). The theme ran through various papers in his Wicksell period. He stressed the need for some understanding of history and of the history of political economy. (On his views, see esp. his contributions to Samuelson [1969*c*]; "The State of Theory in Relation to the Empirical Analysis" [1970*e*]; and "The State of Theory" [1974*a*].)

Of course, he might be accused of not having followed his own precepts. Some of his early writings were not only mathematical by the standards of those days but also overelaborate, and other articles were substantially verbal or geometric mathematics of which the same criticisms could be made. Nevertheless, he always showed an awareness of the need for relevance. This is clear from the preface to *International Trade and Economic Growth* and runs through many (though not all!) of the papers collected in *Aspects of the Theory of Tariffs*. But relevance and simplicity were not his only concerns. In his early days, he had an iconoclastic desire to show that nothing simple said by a British (or other) economist could possibly be true. At times he enjoyed the pleasures of mental gymnastics. In addition, he wanted to show that he could understand and practice complex theory if he really chose.

Finally, something must be said about Harry's broader role in the development of international economics. While this survey has been devoted to Harry's own writings, his contributions to the field go well beyond the work published under his name. His interaction with a vast number of economists is well known. I think he must be given some part of the credit for numerous important articles published by others over a period of more than 20 years. If the list of acknowledgments in a paper contains his name, it can safely be assumed that he made a difference to the paper. Perhaps he encouraged the author initially, perhaps he directed the author's attention to some other work not yet published of which the author should have been aware, or perhaps he suggested redrafting, following up new lines, and so on. Above all, he had the remarkable ability to guide numerous authors in fruitful directions. He was able to see a contribution—however apparently narrow—in the perspective of the whole field and in the light of the scientific development (as he would put it) of the subject.

It can be assumed that he had a significant role in most or all articles in international economics published in the *Review of Economic Studies* and in the *Journal of Political Economy* while he was an editor. He had a flair for seeing which articles were likely to become important in the field. A list of some articles in international trade theory published in the *Review of Economic Studies* or the *Journal of Political Economy* under the Johnson regimes will perhaps make the point: Jones on factor proportions theory (1956), Lipsey and Lancaster on second best (1956–57), Bhagwati on immiserizing growth (1958), Kemp and Jones on variable labor supply (1962), Bhagwati and Ramaswami on domestic distortions (1963), Balassa (1965) and Corden (1966) on effective protection, Cooper and Massell on customs union theory (1965), Kenen on nature, capital, and trade (1965), and Jones on the general equilibrium two-sector model (1965).

His interaction was certainly a two-way process. He kept his eyes and ears open for the latest idea, whether at a seminar or in a draft paper, and if it fitted into his own framework of thinking, he would not hesitate to use it and publish it even before the original author had published it—but always with full acknowledgments, and in fact providing advance publicity for the other's forthcoming paper. He played a key communication role in the field, ensuring that the paper that X was drafting took into account the unpublished ideas of Y in some faraway place. The greatest crime in his eyes was to fail to acknowledge adequately. He could forgive a lack of adequate credits to Samuelson or Hicks but not to the seminal writings of an obscure economist in Australia, Burma, or Canada

Harry was thus "a builder of intellectual
(1978) has correctly stressed. He was a broke
some of the middleman's value added. His wo
striking. It is usual for great economists at a
Chicago or the two Cambridges—to interact w
students, colleagues, or visitors, whether pres
build up "schools." But Harry's circle was worldwide
institution bound. In this respect, as in so many others, he was the complete internationalist. He was the enemy of academic as of other forms of provincialism.

References

Balassa, Bela. "Tariff Protection in Industrial Countries: An Evaluation." *J.P.E.* 78 (December 1965): 573–94.

Barber, Clarence L. "Canadian Tariff Policy." *Canadian J. Econ. and Polit. Sci.* 21 (November 1955): 513–30.

Basevi, Giorgio. "The United States Tariff Structure: Estimates of Effective Rates of Protection of United States Industries and Industrial Labor." *Rev. Econ. and Statis.* 48 (May 1966): 147–60.

Bhagwati, Jagdish N. "Immiserizing Growth: A Geometrical Note." *Rev. Econ. Studies* 25 (June 1958): 201–5.

———. "Protection, Real Wages and Real Incomes." *Econ. J.* 69 (December 1959): 733–48.

Bhagwati, Jagdish N., and Johnson, Harry G. "Notes on Some Controversies in the Theory of International Trade." *Econ. J.* 70 (March 1960): 74–93.

Bhagwati, Jagdish N., and Ramaswami, Vangal K. "Domestic Distortions, Tariffs and the Theory of Optimum Subsidy." *J.P.E.* 71 (February 1963): 44–50.

Branson, William H. "Comment." In Ohlin et al., eds., 1977.

Caves, Richard E. "International Corporations: The Industrial Economics of Foreign Investment." *Economica* 38 (February 1971): 1–27.

Cooper, C. A., and Massell, B. F. "Toward a General Theory of Customs Unions for Developing Countries." *J.P.E.* 73 (October 1965): 461–76.

Corden, W. M. "The Calculation of the Cost of Protection." *Econ. Rec.* 33 (April 1957): 29–51.

———. "The Structure of a Tariff System and the Effective Protective Rate." *J.P.E.* 74 (June 1966): 221–37.

Courchene, Thomas J. "Harry Johnson: Macroeconomist." *Canadian J. Econ.* 11 (suppl.; November 1978): S11–S33.

Dales, J. H. *The Protective Tariff in Canada's Economic Development.* Toronto: Univ. Toronto Press, 1966.

Findlay, Ronald, and Grubert, Herbert G. "Factor Intensities, Technological Progress and the Terms of Trade." *Oxford Econ. Papers* 11 (February 1959): 111–21.

Gamir, Luis. "The Calculation of Effective Rates of Protection in Spain." In *Effective Tariff Protection*, edited by Herbert G. Grubel and Harry G. Johnson. Geneva: General Agreement on Tariffs and Trade, 1971.

Hagen, Everett E. "An Economic Justification of Protectionism." *Q.J.E.* 72 (November 1958): 496–514.

Harrod, Roy F. "Factor-Price Relations under Free Trade." *Econ. J.* 68 (June 1958): 245–55.

Hicks, John R. "An Inaugural Lecture." *Oxford Econ. Papers* 5 (June 1953): 117–35.

Johnson, Harry G. "Optimum Welfare and Maximum Revenue Tariffs." *Rev. Econ. Studies* 19, no. 1 (1951–52): 28–35. Renamed "Alternative Optimum Tariff Formulae." In Johnson 1958*b*.

———. "Equilibrium Growth in an International Economy." *Canadian J. Econ. and Polit. Sci.* 19 (November 1953): 478–500.

——— "Increasing Productivity, Income-Price Trends and the Trade Balance." *Econ. J.* 64 (September 1954): 462–85. (*a*)

———. "Optimum Tariffs and Retaliation." *Rev. Econ. Studies* 21, no. 2 (1951): 142–53. (*b*)

———. "Economic Expansion and International Trade." *Manchester School Econ. and Soc. Studies* 23 (May 1955): 95–112.

———. "Factor Endowments, International Trade and Factor Prices." *Manchester School Econ. and Soc. Studies* 25 (September 1957): 270–83.

——— "The Gains from Freer Trade with Europe: An Estimate." *Manchester School Econ. and Soc. Studies* 26 (September 1958): 247–55. (*a*)

———. *International Trade and Economic Growth: Studies in Pure Theory.* London: Allen & Unwin, 1958. (*b*)

——— "Economic Development and International Trade." *Pakistan Econ. J.* 9 (December 1959): 47–71. (*a*)

———. "International Trade, Income Distribution, and the Offer Curve." *Manchester School Econ. and Soc. Studies* 27 (September 1959): 241–60. (*b*)

———. "The Cost of Protection and the Scientific Tariff." *J.P.E.* 68 (August 1960): 327–45. (*a*)

———. "The Economic Theory of Customs Union." *Pakistan Econ. J.* 10 (March 1960): 14–32. (*b*)

———. "Income Distribution, the Offer Curve, and the Effects of Tariffs." *Manchester School Econ. and Soc. Studies* 28 (September 1960): 215–42. (*c*)

———. "Effects of Changes in Comparative Costs as Influenced by Technical Change." *Malayan Econ. Rev.* 6 (October 1961): 1–13.

———. Review of *An Essay on Trade and Transformation* by Staffan B. Linder. *Economica* 31 (February 1964): 86–90. (*a*)

———. "Tariffs and Economic Development: Some Theoretical Issues." *J. Development Studies* 1 (October 1964): 3–30. (*b*)

———. "The Costs of Protection and Self-Sufficiency." *Q.J.E.* 79 (August 1965): 356–72. (*a*)

———. "An Economic Theory of Protectionism, Tariff Bargaining, and the Formation of Customs Unions." *J.P.E.* 73 (June 1965): 256–83. (*b*)

———. "Optimal Trade Intervention in the Presence of Domestic Distortion." In *Trade, Growth and the Balance of Payments: Essays in Honor of Gottfried Haberler*, by Robert E. Baldwin et al. Amsterdam: North-Holland, 1965. (*c*)

———. "The Theory of Tariff Structure, with Special Reference to World Trade and Development." In *Trade and Development*, by Harry G. Johnson and Peter B. Kenen. Geneva: Librairie Droz, 1965. (*d*)

———. *The World Economy at the Crossroads*. Oxford: Clarendon, 1965. (*e*)

———. "Factor Market Distortions and the Shape of the Transformation Curve." *Econometrica* 34 (July 1966): 696–98. (*a*)

———. "A Model of Protection and the Exchange Rate." *Rev. Econ. Studies* 33 (April 1966): 159–63. (*b*)

———. *Economic Policies toward Less Developed Countries*. Washington: Brookings Inst., 1967. (*a*)

———. "International Trade Theory and M
Theory." In *Monopolistic Competition Theory*: St
Robert E. Kuenne. New York: Wiley, 1967. (*b*)

———. "The Possibility of Factor Price Equalization
number Factors." *Econ. J.* 34 (August 1967): 28

———. "The Possibility of Income Losses from Incr
Accumulation in the Presence of Tariffs." *Econ. J.*
54. (*d*)

———. "Some Economic Aspects of Brain Drain." *Pakistan Development Rev.* (Autumn 1967): 379–411. (*e*)

———. "Comparative Cost and Commercial Policy Theory for a Developing World Economy." Wicksell Lecture. Stockholm: Almqvist & Wiksell, 1968. (*a*)

———. "The Gain from Exploiting Monopoly or Monopsony Power in International Trade." *Economica* 35 (May 1968): 151–56. (*b*)

———. "The Implications of Free or Freer Trade for the Harmonization of Other Policies." In *Harmonization of National Economic Policies under Free Trade*, by Harry G. Johnson, Paul Wonnacott, and Hirofumi Shibata. Toronto: Univ. Toronto Press, 1968. (*c*)

———. "International Trade: Theory." In *International Encyclopedia of the Social Sciences*, vol. 8, edited by David L. Sills. New York: Macmillan, 1968. (*d*)

———. "The Standard Theory of Tariffs." *Canadian J. Econ.* 2 (August 1969): 333–52. (*a*)

———. "The Theory of Effective Protection and Preferences." *Economica* 36 (May 1969): 119–38. (*b*)

———. "The Theory of International Trade." In *International Economic Relations: Proceedings of the Third Congress of the International Economic Association*, edited by Paul A. Samuelson. New York: Macmillan, 1969. (*c*)

———. "The Efficiency and Welfare Implications of the International Corporation." In *The International Corporation: A Symposium*, edited by Charles P. Kindleberger. Cambridge, Mass.: MIT Press, 1970. (*a*)

———. "A New View of the Infant Industry Argument." In *Studies in International Economics: Monash Conference Papers*, edited by I. A. McDougall and Richard H. Snape. Amsterdam: North-Holland, 1970. (*b*)

———. "A Note on Distortions and the Rate of Growth of an Open Economy." *Econ. J.* 80 (December 1970): 990–92. (*c*)

———. "On Factor Price Equalization When Commodities Outnumber Factors: A Comment." *Economica* 37 (February 1970): 89–90. (*d*)

———. "The State of Theory in Relation to the Empirical Analysis." In *The Technological Factor in International Trade*, edited by Raymond Vernon. New York: Columbia Univ. Press (for N.B.E.R.), 1970. (*e*)

———. *Aspects of the Theory of Tariffs.* London: Allen & Unwin, 1971. (*a*)

———. "Effective Protection and General Equilibrium Theory." In Johnson 1971*a*. (*b*)

———. "The Theory of Trade and Growth: A Diagrammatic Analysis." In *Trade, Balance of Payments, and Growth: Papers in International Economics in Honor of Charles P. Kindleberger*, edited by Jagdish N. Bhagwati et al. Amsterdam: North-Holland, 1971. (*c*).

———. "Trade and Growth: A Geometrical Exposition." *J. Internat. Econ.* 1 (January/March 1971): 83–101. (*d*)

——— *The Two-Sector Model of General Equilibrium.* London: Allen & Unwin, 1971 (*e*)

———. "Commercial Policy and Industrialization." *Economica* 39 (August 1972): 264–75. (*a*)

———. "Direct Foreign Investment: Survey of the Issues." In *Direct Foreign Investment in Asia and the Pacific*, edited by Peter Drysdale Canberra: Australian National Univ. Press, 1972. (*b*)

———. "The State of Theory." *A.E.R. Papers and Proc.* 64 (May 1974): 323–24. (*a*)

———. "Trade-diverting Customs Unions: A Comment." *Econ. J.* 84 (September 1974): 618–21. (*b*)

——— "A Note on Welfare-increasing Trade Diversion." *Canadian J. Econ.* 8 (February 1975): 117–23.

———. "A Note on Factor-Income Distribution and the Factor-Owner Production Block." *Manchester School Econ. and Soc. Studies* 44 (December 1976): 303–8. (*a*)

———. "Trade Negotiations and the New International Monetary System." Commercial Policies Issues no. 1, Trade Policy Res. Centre. Leiden: Sijthoff, 1976. (*b*)

———. "A Note on the Geometry of the Two-by-Two General Equilibrium Model." *Economica* 44 (February 1977): 71–75. (*a*)

———. "Technology, Technical Progress and the International Allocation of Economic Activity." In Ohlin et al., eds., 1977. (*b*)

Johnson, Harry G., and Bhagwati, Jagdish N. "A Generalized Theory of the Effects of Tariffs on the Terms of Trade." *Oxford Econ. Papers*, n.s. 13 (October 1961): 225–53.

Johnson, Harry G., and Grubel, Herbert G. "Nominal Tariffs, Indirect Taxes and Effective Rates of Protection: The Common Market Countries 1959." *Econ. J.* 77 (December 1967): 761–76.

Johnson, Harry G., and Krauss, Melvyn B. *General Equilibrium Analysis A Microeconomic Text.* London: Allen & Unwin, 1974.

Johnson, Harry G.; Krauss, Melvyn B.; and Skouras, Thanos. "On the Shape and Location of the Production Possibility Curve." *Economica* 40 (August 1973): 305–10.

Johnson, Harry G., and Mieszkowski, Peter. "The Effects of Unionization on the Distribution of Income: A General Equilibrium Approach." *Q.J.E.* 84 (November 1970): 539–61.

Jones, Ronald W. "Factor Proportions and the Heckscher-Ohlin Theorem." *Rev. Econ. Studies* 24 (October 1956): 1–10

———. "The Structure of Simple General Equilibrium Models." *J.P.E.* 73 (December 1965): 557–72.

Kemp, Murray C., and Jones, Ronald W. "Variable Labor Supply and the Theory of International Trade." *J.P.E.* 70 (February 1962): 30–36.

Kenen, Peter B. "Nature, Capital, and Trade." *J.P.E.* 73 (October 1965): 437–60.

Lerner, Abba P. "The Symmetry between Import and Export Taxes" *Economica* 3 (August 1936): 306–13.

Lipsey, Richard G. "Harry Johnson's Contributions to the Pure Theory of International Trade." *Canadian J. Econ.* 11 (suppl.; November 1978): S34–S54.

Lipsey, Richard G., and Lancaster, Kelvin. "The General Theory of Second Best." *Rev. Econ Studies* 24, no. 1 (1956–57): 11–32

Lizondo, José S; Johnson, Harry G.; and Yeh, Yeong-Her "Factor Intensities and the Shape of the Production Possibility Curve" *Economica* 48 (May 1981): 199–202.

Meade, James E. *Trade and Welfare.* Oxford: Oxford Univ. Press, 1955

———. *The Stationary Economy.* London. Allen & Unwin 1967

Metzler, Lloyd A. "Tariffs, the Terms of Trade
National Income." *J.P.E.* 57 (February 1949)

Ohlin, Bertil; Hesselborn, Per-Ove; and Wijkman
national Allocation of Economic Activity Proceedings
don: Macmillan, 1977

Oniki, Hajime, and Uzawa, Hirofumi. "Patterns of
Dynamic Model of International Trade." *Re*
1965): 15–38.

Pearce, I. F. "The Factor Price Equalisation Myth." *Re*
(1951–52): 111–20

Rybczynski, T. M. "Factor Endowment and Relative Commodity Prices *Economica* 22 (November 1955): 336–41.

Samuelson, Paul A "International Factor-Price Equalisation Once Again" *Econ. J.* 59 (June 1949): 181–97.

Scitovsky, Tibor. "A Reconsideration of the Theory of Tariffs" *Rev Econ Studies* 9, no. 2 (1942): 89–110.

Snape, Richard H. "A New View of the Infant Industry Argument Comment." In *Studies in International Economics*, edited by I A McDougall and Richard H. Snape. Amsterdam: North-Holland, 1970

Viner, Jacob. *The Customs Union Issue* New York: Carnegie Endowment Internat. Peace, 1950.

Young, J. H. *Canadian Commercial Policy* Ottawa Royal Commission Canada's Econ. Prospects, 1957

Harry Johnson as a Macroeconomist

David Laidler

University of Western Ontario

I. Introduction

Harry Johnson's reputation as a macroeconomist rests at least as much on his contributions to policy debates and on his widely read survey articles as on his original research. If one looks at his written work from the narrow vantage point of North America, it is easy to underestimate its importance. Much of what his later writings contain was without a doubt "well known at Chicago" when it was published. However, Harry's capacity for absorbing "oral tradition" was matched by a considerable ability to contribute to it, and it is dangerous to infer a lack of originality from this fact. In any event, in macroeconomics, his chosen audience more often than not consisted of research workers, teachers, and students outside the major centers, not to mention policymakers and "interested laymen." His ability to synthesize and expound the issues being tackled and the results being attained in all areas of the field ensured that their significance was grasped by the profession at large far more rapidly and effectively than otherwise would have been the case. For this reason alone he is a figure of considerable importance in the recent history of macroeconomics.

A number of questions naturally arise. What kind of macroeconomics did he expound? How did it evolve over the years? How much did he himself contribute to the body of knowledge that he expounded? How much did he take from others? And why did the evolution of his thought take the path that it did? Virtually every

---

I am grateful to Victoria Chick, Dennis Coppock, Tom Courchene, Clark Leith, Marcus Miller, Michael Parkin, Don Patinkin, Franco Spinelli, and George Zis for comments on an earlier draft which was presented at the Money Study Group University of Manchester Conference, held in memory of Harry G. Johnson, April 10–11, 1978.

[*Journal of Political Economy*, 1984, vol. 92, no. 4]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9204-0001$01.50

reader of this paper will already have his own views on the answers to the foregoing questions, and it would be as well to outline mine at the outset. Harry's macroeconomics evolved in a direction that led him to pay increasing attention to monetary factors, but he never became a "monetarist" as the term is understood in North America.[1] His originality in this area was not of the very highest order, but by any reasonable standard it was considerable: none of his papers makes a first approach to a new problem, but a number of them represent important contributions to bodies of literature opened up by others. Moreover, his survey papers and policy writings show frequent and highly original insights which he did not always systematically follow up himself. Finally the evolution of Harry Johnson's macroeconomics had very little indeed to do with the fact that he held a chair at the University of Chicago. The evidence of his published work shows that his views stemmed naturally from his perceptions of what the important issues of macroeconomic policy were and from his bringing to bear upon those issues the body of theory and empirical evidence that the published macroeconomic literature provided—including, of course, contributions originating in Chicago.

II. The Cambridge Keynesian

Harry began as what he himself termed a "Cambridge Keynesian" (see McKinnon, ed., 1976, p. 298). In the early 1950s, like many people working on macro and monetary problems in British universities, he subscribed to two only loosely connected bodies of macroeconomics, rather than one: pure monetary theory
about monetary policy as it was actually cond
worked. The former cast only a limited ligh
commentary on Sir Dennis Robertson in the *R*
(1951–52) is purely about the logic of an ec
any clear-cut reference, it is obvious that the
largely Modigliani's (1944) *Econometrica* paper

---

[1] The evolution of Harry's macroeconomics was of course but one aspect of his evolution as a political economist. Over the years he undoubtedly moved from "left" to "right." This was not because he became increasingly confident in the powers of market forces to solve social problems, but rather because he became increasingly pessimistic about the willingness and ability of governments and bureaucracies to solve those same problems. He said as much in the closing passages of the preface to the 1977 edition of his *Canadian Quandary* (1963*b*). It is far beyond the scope of this essay to go into all of these matters. However, let me assert that just as his macroeconomics evolved as he subjected economic analysis to empirical evidence, so I am sure that similar forces governed his broader development. The contrary view, proposed by Bhagwati (1977) that Harry's views changed as a result of "arguing with an ideologue," will no doubt appeal to anyone too lazy to engage in a similar exercise on his own account

from this essay whether, for example, he believed the fixed or flexible price version of the IS-LM model to be the more relevant, or whether he believed the liquidity trap doctrine to be important or unimportant as a practical matter. To Harry in the early 1950s it was the job of economic theory to derive the different implications of different factual assumptions, and that was all. Having just lambasted James Meade (Johnson 1951*b*) for pretending that theory could do more than that, and for slipping his own liberal (in the old-fashioned sense) ideology into what purported to be a theoretical discussion, Harry was understandably careful indeed about maintaining the purity of his own theoretical writing.[2] We have to turn to his articles on policy issues to find out what he thought was of practical significance, as opposed to what he knew to be logically true, in macroeconomics.

The 1950s in Britain saw the revival of monetary policy after a long period of pegged interest rates and direct government control over all aspects of credit allocation. They were also years of creeping inflation. Both circumstances led to substantial academic discussion—and Harry was deeply involved in British debates. His contributions display a considerable uniformity of view until 1956. In 1951, on the very eve of the revival of monetary policy, he published, in the *Economic Journal*, "Some Implications of Secular Changes in Bank Assets and Liabilities in Great Britain" (1951*a*). He noted the immense growth in the money supply and the public debt since the 1930s, and because "overcoming inflationary pressure" rather than "chronic defective demand" was likely to be the major policy problem for the future, he expected those "secular banking trends . . . to complicate rather than to ease the problems of economic policy." Although he also looked for "greater cyclical stability of the money supply" to ease "the problem of contra-cyclical policy," he hastened to note, "Neither of the effects just discussed is one to which any great quantitative significance can reasonably be attached" (1951*a*, p. 556). On the contrary, the significant implication of the facts he described was that "overall economic policy will rely on budgetary measures, buttressed with some degree of direct controls, rather than on monetary measures" (p. 555). This was partly because the use of interest rate variations "to control private borrowing would have significant repercussions on the cost of the public debt" but also because of "the decline which has occurred in the economic significance of the rates of interest which the joint-stock banks directly control," not to mention "the generally accepted view . . . that the interest-elasticity of investment is rather low" (p. 554).

---

[2] In later years, Harry was the first to admit that the review in question was unfair Nevertheless it illustrates an important trait in his work. Though usually the most generous of critics, generosity vanished immediately when, rightly or wrongly, he thought that economics was being pressed into the service of either ideology or personal ambition.

Experience during the early years of the "revival" of monetary policy did not change Harry's view about how policy in fact worked, but he did become more forthcoming about how he thought it ought to work. "The revival of quantitative credit control [i.e., a variable interest rate policy] [has] prepared the way for a genuine return to quantitative [i.e., market-oriented] methods, should the government in future desire it," but the large volume of floating debt and the "British philosophy of government intervention" (1953, p. 26) made that unlikely. Presumably he was not unhappy about that, for he had earlier argued that "the case for the rationing of credit by price . . . is largely irrelevant . . . and open to the objection that the ability to pay high interest rates is not necessarily a proof of the superior social desirability of the project to be financed." He viewed credit rationing by the banks as opposed to government with suspicion on the grounds "that there is no guarantee that the banks will ration out a limited amount of credit in accordance with the national interest" (1952, p. 129).[3]

As late as 1956, to judge from his policy writings, Harry's opinions were still much the same. In a major *Three Banks Review* article, widely read at the time (1956*b*), he identified the view that the quantity of money is a key policy variable with the view that the total volume of bank credit, as opposed to advances, is the central factor. Having thus misunderstood the position, he dismissed it on the grounds that there were many substitute sources of credit for the types of expenditure financed by bank purchases of securities, but
investment financed by advances. Hence the
advances were the important policy variables
"The influence of monetary policy, and the
seem to rest chiefly on its effectiveness as a
holding, and possibly also on its influence
movements; its long-run influence on fixed
Given these views, it is hardly surprising that he con
etary policy is unlikely to be able to do much towards solving the major long-run monetary problem of the British economy, the tendency for British wages and prices to increase more rapidly than those of Britain's international competitors. The most that monetary policy could hope to do would be to prevent the problem from being

---

[3] The reader should note that the comment on the "ability to pay high interest comes from a footnote. In Hutchison (1968) there is a useful account of the contents of the entire symposium, in which Harry (who gives the impression of having been a very serious young man indeed—not yet 30) came in for more than a little leg-pulling from Sir Dennis Robertson and Victor Morgan

periodically aggravated from over-full employment arising from inventory booms" (1956*b*, pp. 18–19).

It was not just because monetary policy was not of any importance in determining aggregate demand that it could not be deployed against inflation. "Substantial unemployment would be required to prevent wages from increasing at an inflationary rate," so neither monetary policy nor "alternative methods of controlling aggregate demand offer much prospect of a solution." Wage-price controls and moral suasion would do no good either: "Such methods attempt to solve the economic problem [of inflation] by ignoring the economic forces which produce it." And what were those forces? "A great deal of the inflationary problem of the post-war years can be attributed to the fact that the increase in productivity since the war has been largely absorbed by increased defence expenditure and the adverse shift in the terms of trade. . . . In future, increasing productivity should permit a steady and substantial growth of real income." This would go far toward creating "the necessary condition for a workable 'national wage-policy'—the possibility of substantial non-inflationary wage increases" (1956*b*, pp. 19–20). In 1956, then, Harry was a proponent of a rather ill-defined cost-push explanation of inflation based on the failure of real income to rise and looked to the future feasibility of an even more ill-defined national wages policy to deal with the problem.

By 1959 Harry's views had changed markedly. Though he still regarded inflation as a "minor evil" that promised to bring no "major economic catastrophe" (hyperinflation or mass unemployment caused by balance of payments crisis as, e.g., Lionel Robbins and Frank Paish tended to argue), Harry nevertheless recognized that it causes "a host of unfairnesses—some cumulative and some temporary, but all irritating—between different members of the community, and makes more complicated and frustrating the tasks of economic calculation for all concerned, including the Government itself."[1] He therefore believed that "the country can probably now afford to trade some loss of output and employment for greater stability of prices" (1958*c*, p. 151). Because he was "inclined to agree . . . that the predominant factor [causing inflation] has been high demand" (p. 149), he also agreed that ". . . demand restriction offers a more effective method of stopping wage inflation than national wages policy. But I would prefer to put the point another way: that a national wages policy is unlikely to be either acceptable or workable unless backed up by adequate control of effective demand" (p. 152).

---

[1] Harry's views concerning mild inflation in developed countries never changed from those expressed here. Indeed, he presented empirical evidence (1963*c*) that, if anything, mild inflation was conducive to economic growth.

Now this shift to a "demand pull" view of inflation is only one aspect of a change in view that becomes evident in Harry's published work after 1956. He also began to put much more emphasis on the quantity of money when discussing policy problems. In 1957 he still believed in the efficacy of direct credit controls "The real bite of the credit squeeze has been achieved . . . through directives relating to bank advances . . . [and] through hire purchase controls," and he urged the Radcliffe Committee to consider whether such "methods could usefully be extended to include a closer supervision over building society lending policies" (1957*a*, pp. 342–43). However, the main burden of the note from which this quotation is taken was to draw attention to the "damning fact" that "no consolidated money supply statement is published in this country" (p 343). The need to improve published statistics was the theme of Harry's advice to the Radcliffe Committee (1960), and by the end of 1958 he had constructed his own annual series for five monetary aggregates for the period 1930–57 (see 1959); the fact that he was willing to engage in that kind of labor, even with the aid of a research assistant, is evidence enough of the importance that he by then attached to the quantity of money. He saw the explanation for the "shortcomings of existing monetary statistics" as lying "partly in the fact that . . . though lip-service is often paid to the idea that monetary policy seeks to control the quantity of money, such control is actually conceived of in terms of its effects on the availability of money, as reflected particularly in conditions in the discount market and the lending activities of the Clearing Banks, partly in the fact that the quantity theory of money, which would provide the maximum incentive for the perfection of monetary statistics, has been so completely discredited, in the eyes of most economists, by its Keynesian rival" (1959, pp 1–2)

The Keynesian rival in question was Modigliani-Hansen IS-LM model that had un tary on Robertson and to the exposition of w tary Theory and Keynesian Economics devoted—and the exogeneous money wage version model at that. However, by 1958 the Pigou effect, which in 1952 had been dealt with in one cryptic parenthetical sentence (see 1951–52, p 103), was presented as reducing that Keynesian theory to "a special case of . . . the neoclassical theory," albeit one that "started from an empirically relevant special assumption, derived some important meaningful results from it, and provided an approach which has since proved its usefulness for a wide range of problems" (1958*a*, p. 240)

For Harry to criticize received Keynesian doctrine in these purely theoretical terms was not new in 1958. He had presented much the same argument at a 1954 conference (see 1957*b*). However, by 1958 he was expressing doubts about that doctrine's practical relevance too. He was now arguing that the Keynesian model "has shown definite weaknesses as a theory of prices." Inflation models based on it "while they describe the inflationary process . . . do not explain it very well. The problem of explaining the differing price histories of different countries during and since the war has led to a certain tendency to return to the quantity theory of money. . . . It is clear that the accumulation of money during the war has had an important effect; but I would myself prefer to employ a neo-Keynesian explanation, based on the accumulation of assets relative to income (these assets including both money and government debt) rather than one which stressed the quantity of money *per se*" (1958*a*, pp. 243–44).

Thus, by 1958, before going to Chicago, Harry had become a very different kind of macroeconomist from the Cambridge Keynesian of the early 1950s. He had of course visited the United States in 1955 and had attended workshops in Chicago, but I can find no evidence of a Chicago influence on his views at this stage, at least on issues of money and inflation. Friedman appeared in his writings at this time as a contributor to the literature on the consumption function, not to that on monetary theory or inflation. The quantity theory–inspired work which he cites is A. J. Brown's *Great Inflation* (1955), not the Chicago *Studies* (Friedman 1957), and the historical experience to which he refers in the passage just quoted is also that dealt with by Brown. His argument for the importance of carrying out statistical studies of transactions velocity (1957*b*) sets him firmly apart from the Chicago School, with its emphasis on income velocity, but also suggests the influence of Brown and his associates, particularly in the light of work later carried out under Brown's supervision on transactions velocity (see Welham 1969). Moreover, in emphasizing the potential importance of money and government debt he looked back to matters which he himself had analyzed 7 years earlier (i.e., 1951*a*), though he now assessed the significance of the relevant data very differently. In short, Harry Johnson's first large step away from orthodox Cambridge Keynesianism was taken at Manchester, not at Chicago, as a response to what he perceived to be the inability of received theory to deal with an important body of evidence.[5]

---

[5] Two points need to be made about the foregoing paragraphs. First, the *Pakistan Economic Journal* article, though published in 1958, has its origins in a lecture delivered in the summer of 1956. I am unable to comment on how close the final published text of the article is to that of the original lecture. Also, it should be noted that there was in no sense a "sudden conversion" of Harry Johnson to the quantity theory of money as a result of reading Brown's *Great Inflation* (1955). He reviewed the book without any great enthusiasm in the *Economic Journal* (1956*a*). Finally, note that the absence of any reference to the Chicago *Studies* in Harry's work up to 1959 may simply reflect the fact that the book was not available in England. Reviews of it did not appear, as far as I have been able to discover, in any British journal before 1959.

III. The Chicago Years

The years 1959–66, when Harry was full-time at Chicago, saw the writing of his best-known and most highly regarded contributions to the macroeconomics literature. The first of these, written during his first year there, was his lecture (1961) on the twenty-fifth anniversary of the publication of the *General Theory*. This completed the break with Cambridge Keynesianism that had begun at Manchester.

The contribution of the *General Theory* as presented in that lecture was only loosely related to the IS-LM model which Harry had hitherto referred to as "Keynesian." Keynes's "central theoretical contribution [was] that in a monetary economy the stability of employment in the face of changes in aggregate demand for output depends on the uncertain monetary effects of changes in money wage levels, which changes may themselves be slow" (1961, p. 3). " 'Unemployment equilibrium' has to be reinterpreted as a disequilibrium situation in which the dynamic adjustment is proceeding very slowly. . . . [This is] a fair modern translation of Keynes's [Marshallian] short-period equilibrium technique" (p. 13).[6] There were weaknesses in the *General Theory*: Keynes "overgeneralized a particularly bad depression" to give the impression that "large-scale unemployment is the typical situation of an advanced capitalist economy," and this was important, given his view that "traditional quantity theory becomes relevant under full employment conditions" (p. 13). The lack of attention to "the influence of price expectations on the assets demand for money, and the associated necessity of distinguishing between real and nominal interest and between fixed-interest-b [illegible] equities" (p. 9) and the neglect of the influenc [illegible] tion were both special cases of the generally [illegible] paid in the *General Theory* to problems of capit [illegible] ever, much harsher criticism was reserved for [illegible] hands "a theory in which money is importa [illegible] theory that money is unimportant" (p. 15). Th [illegible] problem of inflation "concentrate on the mechanism [illegible] causation of inflation; and . . . virtually assume away the possibility of controlling inflation by monetary means" (p. 16) Their bias against according an important role to money had caused the Keynesian approach to inflation "to degenerate into a confused and often obstructive eclecticism" (p. 14).

---

[6] Harry acknowledges the work of Clower, Leontief, and Patinkin in developing this idea He does not give references to specific papers

[7] The reader will note that Leijonhufvud was later to disagree with this judgment Note though that there is no disagreement between Johnson's 1961 lecture and Leijonhufvud's (1968) book on the importance of capital theory for Keynesian economics Their sole disagreement concerned the extent to which Keynes had recognized and coped with this problem himself

In short, his rereading of the *General Theory* caused Harry to make his own distinction between "Keynesian economics" and the "economics of Keynes" and to take the latter as the basis of his own further work in macroeconomics. In the light of this, it is hardly surprising that, though nobody understood the debates about the classical dichotomy and the neutrality of money better than Harry—for which several generations of graduate student readers of his (1962) survey have reason to be grateful—and though nobody more enjoyed wrestling with logical intricacies (Harry always liked puzzles), he recognized from the outset that the debate had little to do with Keynes's attack on the separation of monetary theory and value theory. "The argument . . . has been concerned throughout with a monetary economy characterized by minimal uncertainty, whereas Keynes was concerned with a highly uncertain world . . ." (1962, p. 17). He reiterated this view in a 1963 survey, concluding, "I do not think that this particular controversy is of great relevance to practical work on monetary policy or monetary analysis, or even I think to monetary theory" (1963*d*, p. 84), although he explicitly linked this conclusion to the absence of economic growth from the models in question rather than to the absence of an analysis of uncertainty and expectations on which he based the opening of his critique.[8]

We have already noted that, much as he admired the book, for Harry the great weakness in the *General Theory* was a neglect of capital theory, which resulted from Keynes's use of Marshallian short-period equilibrium to deal with problems of dynamic disequilibrium. Consequently, he valued highly work directed toward making good this deficiency. "Friedman's application to monetary theory of the basic principle of capital theory—that income is the yield on capital, and capital the present value of income—is probably the most important development in monetary theory since Keynes . . ." (1962, p. 33). He praised the contributions of Friedman, Spiro, and Modigliani, Ando, and Brumberg to the consumption function literature for the same reasons, as he praised the contributions to the analysis of what we would now term the "transmission mechanism" that "has been emerging in the past few years, from both 'Keynesian' and 'quantity' theorists, as an outgrowth of the formulation of monetary theory as part of a general theory of asset holding" (1962, p. 50). Again Friedman was cited, this time in company with Tobin, Cagan, and particularly Karl Brunner, from whose work there was a lengthy quotation, as from one who had made an important contribution. But Friedman's contribution in all these cases was seen as being to a theoretical tradition started by Keynes. From the outset Harry was more critical of Friedman's empirical work, particularly of his attempts there to differentiate his approach from the Keynesian tradition as Harry by that time had come to conceive of it.

---

[8] Harry did qualify the foregoing judgment in one respect: He thought the Gurley-Shaw inside money concept (1960) was important as a theoretical matter because of its implications for the Pigou effect. In the light of the outcome of the Pesek-Saving debate, his judgment here was of course erroneous. If we have interest-bearing money, it is not net wealth, but the LM curve is going to be vertical (or at least steep) so there is no need for a Pigou effect. If there is room for substantial variation in the opportunity cost of holding money, on the other hand, it must be because money does not bear a competitively determined rate of return and hence is, on the margin, net wealth.

When he wrote his 1962 survey, Harry had not grasped the interrelationship between Friedman's views on the complexity of the transmission mechanism, the methodology of the famous Friedman-Meiselman study (1963), and the significance of Friedman's (1959) inability to find any well-determined or quantitatively significant relationship between the demand for money and the rate of interest.[9] In 1963, however, Harry pointed to "a certain suspicious appropriateness in the Friedman treatment of the cyclical behaviour of velocity" (1963*d*, p. 102). He returned to this theme more forcefully in his (1965) *Economic Journal* review article of Friedman and Schwartz's *Monetary History* (1963), arguing that Friedman had deliberately downplayed the interest sensitivity of velocity because, among other things, "to admit interest rates into the demand function for money is to accept the Keynesian Revolution and Keynes' attack on the quantity theory" (Johnson 1965, p. 396). In terms of Harry's conception of the nature of Keynes's contribution to economics he was surely right. The liquidity trap doctrine played a minimal role in that conception both because "speculation will take the form of mov securities of different types rather than betw (1961, pp. 8–9), and because the available reduced the existence of a liquidity trap in function to a purely semantic issue (see 1962

Though Harry was not the only reviewer interest elasticity of the demand for money- bin (1965*a*) also made much of it—it was Harry Friedman (1966) into clarifying his views both on the rates in the velocity function—they did have a role to play and on what he believed to be the essential characteristic of Keynesian monetary theory—the liquidity trap. Thus Friedman declined the invitation to embrace the Keynesian revolution, but in doing so he clarified the issues in such a way that those who had regarded him as an anti-Keynesian, but did not regard the liquidity trap as an important ingredient of Keynesian economics, were able to come to a far better appreciation of the significance of his contribution to the subject than they had before.[10]

---

[9] See Johnson (1962) as reprinted in *Essays in Monetary Economics* (*EME*), p. 41. The only way that I can make sense of the first paragraph on this page is to think that Harry was confused between the effects on the transmission mechanism of monetary policy of perfect interest elasticity and perfect interest inelasticity of the demand for money function. The Friedman-Meiselman study (1963) was not of course in its final form when Harry wrote this 1962 survey article. As I recall it, the first draft of that study did not contain an account of the transmission mechanism—though Harry urged the authors to include one—so it is perhaps understandable that he was somewhat muddled on this particular issue at that time.

Even so, critical though he often was of it, Harry always took Chicago empirical work seriously and learned from it. The study of Canadian monetary policy that he and Winder undertook in 1962 was strongly influenced by Friedman and Meiselman (1963), though some aspects of it followed Ando et al.'s (1963) lead. The conclusion he drew from that work (Johnson and Winder 1962), that the effect of monetary policy in the Canadian economy is imprecise, slow, and variable, led him to argue that monetary policy should be used, not for short-run stabilization policy, but for "creating and maintaining a stable long-run monetary environment" (1963*a*, p. 217). This was the *conclusion of Friedman on the one hand and of the Radcliffe Committee* on the other, and in 1963 he was not willing to commit himself either to the Friedman view that stability in the rate of growth of the quantity of money should be the policy target or to the Radcliffe view that stability of interest rates was the key.[11] By the next year he noted evidence produced by Friedman showing that the slackening of U.S. aggregate demand in late 1962 was associated with the previous behavior of the money stock, and he was "inclined to attach a greater importance to monetary policy in . . . sustaining the [subsequent] expansion than the Council [of Economic Advisers] does" (1964, p. 251). Moreover, the effects in question had been the "unintended consequence of a policy intended to be modestly restrictive in the sense of raising the level of interest rates . . . that . . . turned out to be quite expansionary in its effects on the money supply" (p. 251). And for all his criticisms of it, he nevertheless argued that the *Monetary History of the United States* documented "quite conclusively" the presupposition that "the stock of money is the dominant monetary variable in the economy" (1965, p. 389).

---

[10] Friedman's reply to criticism on this score was published in the *Journal of Law and Economics* (1966). His subsequent exchange with Tobin (see Gordon 1971) essentially goes over all this ground again without advancing matters any further. In fairness to Friedman, it should be noted that there is much in the literature on Keynesian economics to justify his view of the central role of the liquidity trap doctrine, particularly the British literature (see, e.g., Hicks 1957 or Kaldor 1970).

[11] This matter is discussed in (1963*a*). On pp. 217–18 of the *EME* reprint, the reader will find a perfect statement of what would now be called the "Poole problem" (cf. Poole 1970). The contribution of Alvin Marty to this paragraph is acknowledged, but it still stands as a particularly good example of the way in which Harry would put highly original formulations of problems into his survey papers and policy writings but not systematically follow them up.

Neither in the papers just cited nor in any other that I have been able to find did Harry come close to arguing that the quantity of money was *all* that matters. He thought that money was more important than did the Council but not "to the point of denying that fiscal policy was an important influence on income and employment" (1964, p. 251); and of course what he regarded as an underemphasis of the importance of real shocks, stemming from downplaying the interest elasticity of velocity, was a key element of his criticism of the *Monetary History*. In 1965 Harry was still by his own lights a Keynesian. Though he was critical of the underlying methodology of the Friedman-Meiselman study (and of the *Monetary History* as well), particularly of the insistence that one simple theory could explain a long run of historical data—"Occam's Razor is a useful principle, but there is no need to cut the throat of empirical research with it" (1965, p. 395)—he accepted its results and did not find anything in them to disturb him: "Any Keynesian can well argue that the 1930's was the period for which the Keynesian theory was designed . . . and that in showing that the Keynesian theory works better for that period Friedman and Meiselman have vindicated the Keynesian Revolution" (1963*d*, pp. 102–3). Nor did he find it difficult to accept the empirical evidence on a stable, albeit interest-sensitive, demand for money function. As he later put it, reiterating a view that he had by then long held, "There is . . . nothing to prevent the absorption
ical evidence of a stable demand function fo
of the Keynesian general equilibrium mo
conditioned Keynesian reflex against the q
the conditioned Keynesian belief that 'mone
least, 'does not matter much')" (1970, pp.

IV. Money, Inflation, and the Monetarist

Though very few people would have called
the mid-1960s (had the label then been m
garded as such in the 1970s and, indeed, often
the label on himself. However, I believe the eviden
work shows that he had gone as far towards monetarism by the mid-1960s as he was ever to go. A further shift towards monetarism after 1965 is an illusion, but why it is such a widely believed-in illusion needs to be explained. Part of the explanation lies in the fact that in 1966 Harry took up his chair at the London School of Economics (LSE) and "conditioned Keynesian reflexes" were more prevalent in Britain than the United States. Kaldor's (1970) celebrated attack on

"The New Monetarism" amounted to a denial of the existence of a stable demand for money function and a reassertion of the unimportance of money; thus, in Britain, a Keynesian of Harry's type found himself labeled a monetarist whether he liked it or not. However, other factors were also important. The problems upon which he worked after 1965 required that he emphasize Friedman's specific contribution to monetary economics more than he had previously, although, as we have seen, he had much earlier recognized and absorbed Friedman's contribution into his own thinking.

As early as his 1961 anniversary lecture, Harry singled out the failure of Keynes to take account of the influence of inflation expectations and systematically to incorporate the distinction between real and nominal interest rates into his work as an important shortcoming of the *General Theory*. He called attention to the importance of Friedman's analysis of this problem in terms of a capital theoretic approach to the demand for money, and a recurring theme of his commentaries on monetary economics thereafter was the weakness of those branches of neo-Keynesian monetary theory that neglected this factor.[12] But of course the influence of inflation expectations is important as a practical matter only if inflation itself is important as a practical matter. As Harry himself remarked, "It is no accident that the appearance of monetarism as a strong intellectual movement has had to wait until the [inflationary] aftermath of the escalation of the war in Viet Nam in 1965" (1971*a*). Nor is it an accident that the aspects of Harry's macroeconomics most readily identifiable as having been learned from Friedman came into prominence in his own work at the same time; the mid-1960s was precisely the time when the analysis of price-level movements began to occupy an increasing amount of his attention. Even so, it should be noted now that one characteristic monetarist belief—that the absence of a long-run inflation/unemployment trade-off is an *empirically* important matter—does not, as far as I have been able to discover, and to my considerable surprise, appear in his writings.

Like any other macroeconomist, Harry was compelled by the circumstances of the late 1960s and early 1970s to pay attention to inflation as a policy problem, but he also at that time carried out important original work in monetary theory. Characteristically, there was a substantial spillover from this work on pure theory into his policy writings, and though the theoretical literature to which he contributed has become unfashionable, his work there still retains interest if only because of this spillover. The literature in question is epitomized in such phrases as "money in a growth model" and "money and welfare."

---

[12] In particular, the work of Tobin and the Yale School on portfolio behavior is systematically criticized for this (see, e.g., Johnson 1970, 1974), but note that Tobin's work on money and economic growth is immune to this criticism

Harry had long been dissatisfied with the literature on the classical dichotomy and neutral money, not only because it was not dealing with the problem in the context of the uncertain world which he regarded as fundamental to Keynesian economics, but also because it abstracted from economic growth. The literature on money and growth meets the second of these objections, but only by completely ignoring the first: that is one reason why it has now fallen out of fashion. Harry's particular contribution to it (1966), in contrast to Tobin's initial work (1965*b*), "stressed the necessity of allowing money a function in a monetary economy, and therefore of attributing to the presence of money an increase in economic welfare" (1970, p. 109). In short, this paper incorporated Friedman's analysis of money "as if" a consumer durable; it showed that because the rate of inflation influenced holdings of real balances it would also influence "real income" and hence the savings rate, the details depending on the way in which savings behavior was modeled. The conclusion established by this literature—and Harry's contribution sets out all the essential results while managing also to be the most readable paper on the subject—is that the full-employment equilibrium properties of a monetary economy are extremely unlikely to be independent of monetary factors (specifically the rate of change of the nominal money supply). This conclusion holds even if all Keynesian problems about uncertainty are assumed away by making inflation, and
that matter, "fully anticipated."[13] Since one
employment" equilibrium that fully anticip
is presumably the "natural" unemployme
question may be relevant to important pol
vance is still open to doubt.[14] In my view
of the upshot of the closely related money
erature, it is likely to remain so.

As is well known, Pesek and Saving (1967) showed
outside money" distinction of Gurley and Shaw—to which Harry devoted considerable space in his early 1960s surveys—stemmed from endowing accounting conventions with spurious economic content. Once we leave a commodity money world, the key distinction—and this only became clear after a somewhat convoluted discussion to which Harry made an important contribution (see 1969*b*, 1969*c*)—is

---

[13] See Johnson (1966). Some of the analysis there presented is further developed in the latter half of "Inside Money, Outside Money" (1969*b*).

[14] I am indebted to Douglas Purvis for some discussions on this point.

between money that bears interest at competitive rates and that which does not because its issue is monopolized. But all the results of the money in growth-models literature and the "schizophrenic policy proposals" that emerged from Friedman's "Optimum Quantity of Money" (1969) essay hinge upon variations in the rate of nominal monetary expansion influencing the opportunity cost of holding money. Thus, as Harry was quicker than anyone else to stress (see 1968*b*), they depend on the assumption of a particular institutional anomaly—namely, that the monetary system is not competitive. Only if "the practical point that the payment of interest on currency holdings is infeasible . . . is taken as being of predominating importance" (1970, pp. 106–7) does the rate of nominal monetary expansion have any long-run significance, and Harry was sufficiently unimpressed with its importance to be willing to abstract from it entirely in his most extensive discussion of these very issues (1968*b*). Otherwise, as far as he was concerned, the welfare analysis of money was significant in two respects: first, as a component of the case for more competition in the banking system, this was an important policy issue in Britain in the late 1960s; and, second, inasmuch as it dealt with the transition from commodity to fiat and credit money systems, it yielded insights into the problems of reforming the international monetary system.[15]

The notion of a fully anticipated inflation rate is fundamental to the literature on money in a growth-model or money-and-welfare literature. The recognition that if agents allow for inflation expectations in their portfolio behavior they are also going to allow for them in their behavior toward money wages and prices underlies the literature on the expectations-augmented Phillips curve. The "accelerationist hypothesis" follows from the proposition that, in the long run, inflation is indeed fully anticipated—that is, it is expected to occur and behavior is completely adapted to the expectation in question. Like everyone else in the 1960s, Harry was slow indeed to incorporate expectations into his treatment of the Phillips curve, and once he had, though he immediately grasped the logic of the accelerationist hypothesis as a piece of economic theory, he never embraced it as a hypothesis relevant to the practical conduct of monetary policy in an inflationary environment.

In his article "A Survey of Theories of Inflation" (1963*e*) Harry provides as good an example as one could find of a man who knows something without knowing that he knows it. He criticized "Keynesian models" that analyze inflation in terms of competition between various social groups for greater shares in national income because, unlike the "quantity theory approach," which he identified with the

---

[15] On the international monetary system see Johnson (1968*a*). On competition in banking see Johnson (1967*b*).

Cagan-Bailey analysis of the inflation tax, they failed to recognize "that the processes of determining wages and prices are fundamentally real processes." Protection against inflationary redistributions "is available to all groups in a freely competitive system, through negotiation of contracts to take account of expected inflation, and is increasingly resorted to as the fact of inflation is recognized" (1963*e*, p. 122). He even pointed out that "there are some serious doubts about the applicability of the Phillips curve to the formulation of economic policy. . . . it may reasonably be doubted whether the curve would continue to hold its shape if an attempt were made by economic policy to pin the economy down to a point on it" (pp. 132–33). Nevertheless, "the Phillips curve appears to be far the most reliable of [the] relationships" describing statistical trade-offs between policy goals (p. 133). It formed the basis for his subsequent discussion of policy trade-offs; and Reuber's pioneering (1964) attempt to quantify, in dollars and cents terms, the trade-off between inflation and unemployment was held up as an example of the type of work which must be further developed so that the curve could be used as a basis "for intelligent policy-making" (1963*e*, p. 141).

It is not difficult to understand why Harry fell into this particular inconsistency. His writings over the years make it clear that he thought the rapid inflations experienced by underdeveloped countries presented a very different set of problems from the mild inflations of the pre-1970s in the developed world (see 1967*a*). The Phillips curve was a potentially interesting tool for [illegible] while the quantity theory approach had prov[illegible] in dealing with the former. "For the mild ty[illegible] the United States and other advanced coun[illegible] theory] approach has not proved nearly so [illegible] rate of price change has not appeared as a [illegible] the quantity of money demanded" (1963*e*, p[illegible] pirical evidence suggested to him that the exp[illegible] unimportant in such a context, there was no [illegible] incorporate it systematically into his thinking about [illegible] Thus even when, in the Introduction to the second edition of *Essays in Monetary Economics* (1969*a*), he noted the inconsistency in question, he argued that it was of more theoretical than practical significance. The importance of incorporating inflation expectations into the Phillips curve ". . . for the theory of macro-economic policy . . . depends on the length of time it takes for changes in expectations generated by experience to begin to affect behaviour significantly; and the empirical evidence is that the lags in adjustment of expectations are sufficiently long for contemporary policy makers safely to disregard them—even though the cumulative effect may be substantial" (p. x).

I have already said that I have found no statement in any of Harry's writings of a belief in the empirical, as opposed to theoretical, relevance of a vertical long-run Phillips curve. However, in his later work he certainly did pay increasing attention to the role of expectations in determining the economy's response to stabilization policy and expressed increasing doubts about the theoretical foundations of the Phillips curve itself. Thus, he argued (1968*b*) that "stabilization operations [themselves might disturb] expectations derived from previous experience" by way of "mechanisms [that] cannot be . . . satisfactorily dealt with by compressing them into the distributed lag structure of the economy" (p. 301). By 1972 he had come to the view that "contrary to the standard assumptions of economic theory, the economic public does not simply respond mechanically . . . to signals reaching it through the blind and impersonal operations of . . . competitive markets. Instead [it] engages in two kinds of political transactions with . . . policy-makers. . . . First, one of its major concerns is to guess how determined the government is about implementing its announced economic policies. . . . Second, the relevant economic public, aware of the sensitivity of government to political pressures, has an incentive to generate such pressures in its own favour" (1972, pp. 14–15).[16] As to the Phillips curve itself, new work "broadening the microeconomic foundations of the relationship in terms of the influence of information and adjustment costs and other real factors" (1972, p. 66) was required, a theme which he reiterated at greater length (1974).

Although he regarded neither the theoretical nor the empirical basis of the relationship to be well established, as an empirical matter he seems nevertheless to have believed in the existence of a long-run trade-off between inflation and unemployment, basing that belief on his reading of the empirical evidence—particularly that of Solow, whose work is cited in a survey (1970), lecture notes (1971*b*), and the DeVries lectures (1972).[17] On the empirical nature of the Phillips curve, Harry's position was thus closer to Tobin's than to Friedman's. It is worth noting that Harry's major reason for predicting the ultimate failure of the monetarist counterrevolution (1971*a*) was rooted in his perception of monetarism's inability to provide a satisfactory theory of unemployment to complement its theory of inflation.[18] However, in Britain in the early 1970s the belief that the excess demand for labor could exert any influence on the inflation rate was enough to get its adherent classified as a monetarist, regardless of his views about the role that inflation expectations, information costs, and such might play in the process, particularly if that view was combined, as it was in Harry's case, with opposition to wage and price controls.

---

[16] In the first of these quotations the reader will recognize the germ of the "rational expectations" idea. The second quotation represents a substantial extension of the notion, one that has not yet been taken up in the analytic literature. Harry's only explicit comments on the hypothesis in question occur in the introduction to his *Selected Essays* (1978), which are notable for his caution over the implications of a thoroughgoing application of the hypothesis to problems of short-run stabilization policy.

[17] Of course he should not have taken Solow's results seriously, because those tests were flawed (see Laidler 1970). However, it is easy to forget that in 1971 virtually all the available empirical evidence seemed to show that there was a long-run inflation/unemployment trade-off. It has only been with the growing sophistication of the modeling of expectations that results favorable to the long-run vertical Phillips curve have been generated. By the time that work was produced, Harry was mainly working on other matters, and only in the introduction to his *Selected Essays* (1978) does there appear any hint that such work might have persuaded him to change his views. However, there is no definite statement to this effect.

Opposition to direct controls on wages and prices was a constant characteristic of Harry's writings on inflation from his very earliest papers in the 1950s to his last ones. He consistently argued that wage and price controls would make the effects of inflation worse without contributing anything to curing the disease. As he put it (1963*e*, pp. 134–35): "Does inflation necessarily lead to monetary collapse? The answer is no, provided that the price system is uncontrolled and prices are allowed to rise freely. . . . Where monetary breakdowns do occur is where rapid inflation is combined with general price control, so that money ceases to be usable for making transactions This is the explanation, for example, of the breakdown of the German currency after World War II. . . . Does inflation cause economic inefficiency? . . . The evidence seems to indicate that inflation does not reduce economic efficiency, at least when the inflation is mild and reasonably steady and prices and wages are not co[illegible] Harry's contribution to Parkin and Sumner (197[illegible] 1977 edition of *The Canadian Quandary* re[illegible] albeit with extra vehemence, born of the [illegible] ing controls as a current policy issue rath[illegible] demic point of view. But let it be noted [illegible] such controls is cited as evidence of "mone[illegible] part, then he had acquired those leanin[illegible] Chicago (see 1956*b*)

Harry's role in developing the monetary approach to balance-of-payments theory is beyond the scope of this paper Some mention must be made of it at this point, however, a misunderstanding of

---

[18] The Ely Lecture to which I refer caused quite a stir when it was delivered, containing as it did attacks on both Keynes's and Friedman's means of propagating their ideas. If anything, that lecture is a contribution to the sociology of knowledge rather than to macroeconomics, and that is why I do not pay more attention to it here. It is ironical that Harry himself was subsequently criticized along identical lines in connection with his work on the monetary approach to balance of payments theory (see Hahn 1977, Coppock 1978).

Harry's own view of that theoretical development has contributed also to the opinion that he was a "monetarist." In 1974 he accepted that "criticisms directed at this approach for its neglect of dynamic disequilibrium behaviour, [and] the processes of international transmission of monetary impulses . . . are certainly justified by the published literature of the approach as it now stands" (1974, p. 224). However he argued that "some at least . . . recognize that reality occupies the area between the closed and the open economy theoretical assumptions, very much as the problem of how the effects of monetary impulses are divided between price and quantity responses defines the grey area requiring intensive empirical research in contemporary closed-economy monetary economics" (p. 224). His own research program at the LSE was "intended to pursue this matter in depth." In short, Harry's published writings on the monetary approach, "monetarist" though their characteristics undoubtedly were, particularly in their reliance on the assumption of full employment and the hypothesis of a stable demand for money function, were meant to set out a theoretical basis for organizing empirical work, not to state his beliefs about the nature of the world.

V. Conclusions

There is no neat and tidy ending to the story of the development of Harry Johnson's macroeconomics. He had always paid more attention than did the majority of contributors to the area to open-economy problems, and by 1969 he concluded, "Perhaps the greatest disservice that Keynes rendered to the development of economics [in Britain] was to develop the theory of macro-economics and money on the assumption of a closed economy. The extension of Keynesian theory to an open economy . . . has been built on the manifestly unsatisfactory assumption of money illusion on the part of wage-earners. . . . Much work remains to be done in developing a monetary economics appropriate to the analytical and policy problems of the British economy" (1970, p. 114). Thereafter the development of what has come to be called "the monetary approach to balance-of-payments theory" occupied an increasing proportion of his time, so that the 1970s saw very little from him that falls within the traditional boundaries of macroeconomics.

His survey article for *Oxford Economic Papers* (1974), already alluded to, and a *Journal of Economic Literature* paper written with Bob Nobay (Johnson and Nobay 1977) are the only major pieces of work. The latter, extending ideas originally set out first briefly in the DeVries lectures and later in a note in *Kredit und Kapital* (1976), is interesting for the parallels that it draws between the pre-Keynesian analysis of short-run monetary equilibrium and recent developments in American monetarism, especially those associated with the work of Lucas, Sargent, and Wallace on rational expectations. However, it lies outside the mainstream of his later work. That, to repeat, concerned the macroeconomics of the open economy; to deal with it would take me beyond the assigned scope of this paper. Let it be stressed, though, that Harry never regarded that work as dealing with a separate field. He viewed it as a natural development of his work on macroeconomics and money in which he integrated Friedman's notion of a stable demand for money function with his own, much earlier, attempts at generalizing balance-of-payments theory (1958*b*). Thus the reader of this essay should weigh his work on the monetary approach to balance-of-payments theory along with the contributions with which this essay has dealt when he assesses Harry's importance as a macroeconomist.

Of course, it is far too early to carry out such an assessment in any final way. What can be done, though, is to sum up the salient features of Harry's macroeconomics as it emerges from the work which has been surveyed in this essay. First and foremost, of course, it is remarkable for its breadth, but it is remarkable also for the consistency of its approach. In particular, at every stage we can see that Harry paid particular attention to empirical evidence in forming his views. It was the inflation of the 1950s that persuaded him to pay more attention to monetary factors; it was his own empirical work and that of Brown and of Friedman and his associates that persuaded him of the particular importance of the quantity of money and of the futility of using monetary policy for short-run stabilization policies; it was empirical evidence drawn from underdeveloped countries that [illegible] of the importance of the peculiarly "Chicago [illegible] tary economics (namely, the role of inflation [illegible] ing the demand for money); and it was em[illegible] him to his views on the importance of infl[illegible] context of the policy problems of developed [illegible] to his respect for empirical evidence is the cons[illegible] searched for policy relevance in his macroeconomics [illegible] the Pigou effect, neutrality, and the classical dichotomy were examined for such relevance and found wanting; the literature on money, welfare, and growth provided him with tools to analyze the reform of domestic banking and the international monetary system; he always regarded unemployment as a more important policy problem for developed countries than inflation, so Friedman's contributions were to him a complement to Keynes's economics, not a substitute, and his LSE research program was designed to investigate the empirical content and policy relevance of the monetary approach to balance of payments analysis.

For Harry Johnson "the purpose of economics as a social science is to arrive at a set of principles for understanding and interpreting the economy that are both scientifically 'robust' and sufficiently simple to be communicable to successive generations of students and policymakers and the general public" (1974, p. 214). No one worked more diligently and consistently to further that purpose than he did.

References

Ando, Albert, Brown, E. Carey, Solow, Robert M.; and Kareken, John. "Lags in Fiscal and Monetary Policy." In *Stabilization Policies*, by the Commission on Money and Credit. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Bhagwati, Jagdish N. "Harry G. Johnson." *J. Internat. Econ.* 7 (August 1977): 221–29.

Brown, Arthur J. *The Great Inflation, 1939–1951*. London: Oxford Univ. Press, 1955.

Coppock, Dennis J. "The Monetary Theory of the Balance of Payments: A Commentary." Mimeographed. Manchester: Univ. Manchester, 1978.

Friedman, Milton, ed. *Studies in the Quantity Theory of Money.* Chicago: Univ. Chicago Press, 1957.

———. "The Demand for Money: Some Theoretical and Empirical Results." *J.P.E.* 67, no. 4 (August 1959): 327–51.

———. "Interest Rates and the Demand for Money." *J. Law and Econ.* 9 (October 1966): 71–85.

———. "The Optimum Quantity of Money." In *The Optimum Quantity of Money, and Other Essays.* London: Macmillan, 1969.

Friedman, Milton, and Meiselman, David. "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1898–1958." In *Stabilization Policies*, by the Commission on Money and Credit. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Friedman, Milton, and Schwartz, Anna J. *A Monetary History of the United States, 1867–1960.* Princeton, N.J.: Princeton Univ. Press (for N.B.E.R.), 1963.

Gordon, Robert J., ed. *Milton Friedman's Monetary Framework: A Debate with His Critics.* Chicago: Univ. Chicago Press, 1974.

Gurley, John G., and Shaw, Edward S. *Money in a Theory of Finance.* Washington: Brookings Inst., 1960.

Hahn, F. H. "The Monetary Approach to the Balance of Payments." *J. Internat. Econ.* 7 (August 1977): 231–49.

Hicks, John R. "A Rehabilitation of 'Classical' Economics?" *Econ. J.* 67 (June 1957): 278–89.

Hutchison, Terence W. *Economics and Economic Policy in Britain, 1946–1966: Some Aspects of Their Interrelations.* London: Allen & Unwin, 1968.

Johnson, Harry G. "Some Implications of Secular Changes in Bank Assets and Liabilities in Great Britain." *Econ. J.* 61 (September 1951): 544–61. (*a*)

———. "The Taxonomic Approach to Economic Policy." *Econ. J.* 61 (December 1951): 812–32. (*b*)

———. "Some Cambridge Controversies in Monetary Theory." *Rev. Econ. Studies* 19, no. 2 (1951–52): 90–104.

———. "The New Monetary Policy and the Problem of Credit Control." *Bull. Oxford Univ. Inst. Statis.* 14 (April/May 1952): 117–31.

———. "Recent Developments in British Monetary Policy." *A.E.R. Papers and Proc.* 43 (May 1953): 19–26.

———. Review of *The Great Inflation, 1939–1951*, by Arthur J. Brown. *Econ. J.* 66 (March 1956): 121–23. (*a*)

———. "The Revival of Monetary Policy in Britain." *Three Banks Rev.* 30 (June 1956): 3–20. (*b*)

——— "Bank Rate Reform and the Improvement of Monetary Statistics." *Bull. Oxford Univ. Inst. Statis.* 19 (November 1957): 341–45. (*a*)

——— "The Determination of the General Level of Wage Rates." In *The Theory of Wage Determination*, edited by John T. Dunlop. London: Macmillan, 1957. (*b*)

———. "Monetary Theory and Keynesian Economics." Lecture no. 5 of series delivered to Pakistan Economics Association. *Pakistan Econ. J.* 8 (June 1958): 56–70. Reprinted in *Monetary Theory. Selected Readings*, edited by Robert Clower. Baltimore: Penguin, 1969. (*a*)

———. "Towards a General Theory of the Balance of Payments." In *International Trade and Economic Growth. Studies in Pure Theory.* London: Allen & Unwin, 1958. (*b*)

———. "Two Schools of Thought on Wage Inflation." *Scottish J. Polit. Econ.* 5 (June 1958): 149–53. (*c*)

———. "British Monetary Statistics." *Economica* 26 (February 1959): 1–17.

———. "Memorandum of Evidence." In *Principal Memoranda of Evidence*, vol. 3, by the Committee on the Working of the Monetary System. London: Her Majesty's Stationery Office, 1960.

———. "The *General Theory* after Twenty-five Years." *A.E.R. Papers and Proc.* 51 (May 1961): 1–17.

———. "Monetary Theory and Policy." *A.E.R.* 52 (June 1962): 335–84. Reprinted in *EME*, 1969*b*.

———. *Alternative Guiding Principles for the Use of Monetary Policy in Canada.* Essays in International Finance no. 44. Princeton, N.J.: Princeton Univ. Press, 1963. Reprinted in *EME*, 1969*a*. (*a*)

———. *The Canadian Quandary: Economic Problems and P* Toronto: McGraw-Hill, 1963. Carleton Library ed., Toronto: M
1977. (*b*)

———. "Objectives, Monetary Standards and P
Statis. 45 (October 1963): 137–47. (*c*)

———. "Recent Developments in Monetary Th
(August 1963): 1–28. Reprinted in *EME*, 196

——— "A Survey of Theories of Inflation." *I*
1963): 29–69. Reprinted in *EME*, 1969*a*. (*c*)

———. "Major Issues in Monetary and Fiscal Policies
(November 1964): 1400–1413. Reprinted in *EME*

——— "A Quantity Theorist's Monetary History of the United States." *Econ. J.* 75 (June 1965): 388–96.

——— "The Neo-Classical One-Sector Growth Model: A Geometrical Exposition and Extension to a Monetary Economy." *Economica* 33 (August 1966): 265–87. Reprinted in *EME*, 1969*a*.

——— "Is Inflation the Inevitable Price of Rapid Development or a Retarding Factor in Economic Growth?" In *Fiscal and Monetary Problems in Developing States*, edited by David Krivine. New York: Praeger, 1967. Reprinted in *EME*, 1969*a*. (*a*)

———. "The Report on Bank Charges. II." *Bankers' Magazine* 204 (August 1967): 64–68. Reprinted in *Readings in British Monetary Economics*, edited by Harry G. Johnson. Oxford: Clarendon, 1972. (*b*)

——— "A Note on Seigniorage and the Social Saving from Substituting Credit for Commodity Money." *Univ. Punjab Economist* 6 (June 1968): 1–8.

Reprinted in *Monetary Problems of the International Economy*, edited by Robert A. Mundell and Alexander K. Swoboda. Chicago: Univ. Chicago Press, 1969. (*a*)

———. "Problems of Efficiency in Monetary Management." *J.P.E.* 76, no. 5 (September/October 1968): 971–90. Reprinted in *Readings in British Monetary Economics*, edited by Harry G. Johnson. Oxford: Clarendon, 1972. (*b*)

———. *Essays in Monetary Economics*. 2d ed. London: Allen & Unwin, 1969. (*a*)

———. "Inside Money, Outside Money, Income, Wealth, and Welfare in Monetary Theory." *J. Money, Credit and Banking* 1 (February 1969): 30–45. (*b*)

———. "Pesek and Saving's Theory of Money and Wealth: A Comment." *J. Money, Credit and Banking* 1 (August 1969): 535–37. (*c*)

——— "Recent Developments in Monetary Theory: A Commentary." In *Money in Britain, 1959–69: The Papers of the Radcliffe Report Ten Years After*, edited by David R. Croome and Harry G. Johnson. London: Oxford Univ. Press, 1970.

———. "The Keynesian Revolution and the Monetarist Counter-Revolution." *A.E.R. Papers and Proc.* 61 (May 1971): 1–14. Reprinted in *Further Essays in Monetary Economics*. London: Allen & Unwin, 1972. (*a*)

———. *Macroeconomics and Monetary Theory*. London: Gray-Mills, 1971. (*b*)

———. *Inflation and the Monetarist Controversy*. DeVries Lectures. Amsterdam. North-Holland, 1972.

——— "Major Issues in Monetary Economics." *Oxford Econ. Papers* 26 (July 1974): 212–25.

———. "Comment on Mayer on Monetarism." *Kredit und Kapital* 9, no. 2 (1976): 145–53.

———. *Selected Essays in Monetary Economics*. London: Allen & Unwin, 1978.

Johnson, Harry G., and Nobay, A. Robert. "Monetarism: A Historic-Theoretic Perspective." *J. Econ. Literature* 15 (June 1977): 470–85.

Johnson, Harry G., and Winder, John W. L. *Lags in the Effects of Monetary Policy in Canada*. Ottawa: Queen's Printer (for Royal Commission on Banking and Finance), 1962.

Kaldor, Nicholas. "The New Monetarism." *Lloyds Bank Rev.* (July 1970), pp. 1–18.

Laidler, David. "Recent Developments in Monetary Theory: Discussion." In *Money in Britain, 1959–69. The Papers of the Radcliffe Report Ten Years After*, edited by David R. Croome and Harry G. Johnson. London: Oxford Univ. Press, 1970

Leijonhufvud, Axel. *On Keynesian Economics and the Economics of Keynes*. London: Oxford Univ. Press, 1968.

McKinnon, Ronald I., ed. *Money and Finance in Economic Growth and Development: Essays in Honor of Edward Shaw* New York and Basel: Dekker, 1976.

Meltzer, Allan H. "Monetary Theory and Monetary History." *Schweizerische Zeitschrift für Volkswirtschaft und Statistik* 101 (December 1965): 404–22.

Modigliani, Franco. "Liquidity Preference and the Theory of Interest and Money." *Econometrica* 12 (January 1944): 45–88.

Parkin, Michael, and Sumner, Michael T., eds. *Incomes Policy and Inflation*. Manchester: Manchester Univ. Press, 1972.

Pesek, Boris P., and Saving, Thomas R. *Money, Wealth, and Economic Theory*. New York: Macmillan, 1967.

Poole, William. "Optimal Choice of Monetary Policy Instruments in a Simple Stochastic Macro Model." *Q.J.E.* 84 (May 1970): 197–216.

Reuber, Grant L. "The Objectives of Canadian Monetary Policy, 1949–61:

Empirical 'Trade-Offs' and the Reaction Function of the Authorities." *J.P.E.* 72, no. 2 (April 1964): 109–32.

Tobin, James. "The Monetary Interpretation of History." *A.E.R.* 55 (June 1965): 464–85. (*a*)

———. "Money and Economic Growth." *Econometrica* 33 (October 1965): 671–84. (*b*)

Welham, P. J. *Monetary Circulation of the United Kingdom: A Statistical Study.* Oxford: Blackwell, 1969.

Harry G. Johnson as a Development Economist

Arnold C. Harberger

University of Chicago

David Wall

University of Sussex

I. Introduction

In making an assessment of Harry Johnson's work in the area of development economics one must recognize (*a*) that as an individual he belongs first and foremost to the whole discipline of economics rather than to any particular branch or field thereof and (*b*) that, certainly in his own view, economic development is not itself a separate field of economics in the same sense as, say, international trade or monetary economics (the areas in which his roots were deepest, his interests strongest, and his contributions most numerous and powerful). That he belongs to all of economics is reflected in the enormous scope and range of his work. His denial of separate "field" status to economic development (a position we also share) had as its counterpart a firm conviction that the discipline of economics had great relevance for all countries, regardless of their stage of development. He preferred the term "development economics" to "economic development," because he found it more compatible with his views that there is really only one science or discipline of economics and that most important policy problems have to be faced by economies at all levels of wealth (or stages of development).

The main task of development economics was, in Johnson's view, simply to apply good economics to the analysis of the economic situa-

We are grateful to Nancy Wall for her help in the preparation of this paper.

[*Journal of Political Economy*, 1984, vol. 92, no. 4]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9204-0007$01.50

tions of one or more of the less developed countries (LDCs) and to the study of particular policy problems facing them. Sometimes his applications were quite straightforward, entailing only modest adaptations (if any) from the way the same task would be done in a typical advanced economy. In this vein, much of Johnson's "advice" to the LDCs took the form of trying to convince them that they were not (either individually or collectively) "special cases" of one or another particular problem. Sometimes, however, without being subject to a different kind of economics, a less developed country could have certain characteristics (e.g., one-crop export agriculture) that led to problems different from those of more advanced countries. At other times questions arose, usually in international forums, concerning the relationship of the LDCs as a group to the principal industrial centers. Dealing with these questions, too, was in Johnson's view a legitimate part of development economics.

We are not interested even in attempting to set what would in any event be arbitrary boundaries to "development economics," nor indeed was Johnson. The main point is quite the opposite—development economics is often just ordinary economics applied to less developed countries, and where it is different the differences are so varied and often so subtle that the idea of a clear borderline is itself absurd. This in turn makes it difficult to say just which of Johnson's many works should or should not be considered as falling within the scope of this review. In spite of the difficulty, we have done our best to cull from his vast bibliography a list of works that seem to us to fit reasonably comfortably under the rubric of development economics as described above.

Apart from a single entry in 1950, "The
International Commodity Agreements on the
ucts," no other met the test until the artic
Development and the Maximum Rate of Gro
the *Malayan Economic Review*. This was but one
series of papers published in that journal, the
view, the *Pakistan Development Review*, the *Philippine*
and a few other similar outlets; in such cases we took the outlet itself to create a presumption (in what might otherwise be a doubtful case) for including the paper in our "development economics" list.

Despite all of the difficulties of classification, we consider our selected list to be revealing in the dimensions both of time and of subject matter. On the time scale, after one article in 1950, there was one in 1957, two in 1958, and one in 1959. There were three each in 1960 and 1961, followed by a hiatus of sorts (one article in 1963), after which Johnson's productivity in the development economics field really burgeoned, with 20 articles in 1964–66 and another 10 in

1967–68. The pace fell off to three or fewer per year from 1969 through 1973, followed by a spate of 18 articles in 1974–77. Clearly, the mid-1960s formed the first major period of concentration, and during this time there were also four books (1965*c*, 1967*b*, 1967*c*, 1968) that we felt merited being classified as dealing with development economics. The second major period was the mid-1970s; here there were an additional three books (1975, 1976*a*, 1977*b*) bearing on this field.

Subject matter shows a significant shift of emphasis through time. Articles that we would classify as dealing principally with customs unions and with tariffs and trade restrictions are significantly concentrated in the 1950–61 period (five articles) and 1963–68 period (six articles). The policies of the developed countries (most particularly the United States) vis-à-vis the Third World enter strongly (seven articles) in the 1963–68 period, while the themes of UNCTAD and the so-called link between international aid and the international monetary system (six articles) come to the fore during the 1970s. Papers dealing with the general area of economic development (10 articles), on the other hand, are scattered quite evenly along the time dimension.

In the sections that follow, we try to convey both the content and the "flavor" of Johnson's work in development economics. Wherever possible we use his own words, for in most cases it was possible to find, among his various writings, one or more versions in which a given argument or line of thought was quite elegantly distilled.

The organization of the remainder of this review is as follows. In Section II we present Johnson's view of the development process. Subsequently, we deal with his writings, first (in Sec. III) concerning economic policy within the developing countries and then (in Sec. IV) concerning policies of the advanced countries that had particular bearing on the LDCs. Section V deals with Johnson's thoughts on the mobility across national boundaries of factors of production (labor and capital) and of technology, while Section VI goes into the whole complex of problems encompassing international arrangements, including both developing and advanced economies. These include international commodity agreements, trade preferences for the products of LDCs, and institutional arrangements such as UNCTAD and the "link" (between the creation of international liquidity on the one hand and the distribution of foreign aid to the LDCs on the other).

Our final section treats the more human aspects of Harry Johnson's connection with development economics—his incessant travels to and labors in developing countries, his personal ties to scores of former students, and his efforts to aid young representatives of modern economics in their struggles to improve the teaching, the application, and the diffusion of economic science throughout the developing world.

II. The Development Process

Starting at least with some early postwar writings, much of the literature on economic development has been characterized by mechanistic assumptions of various types. Johnson was at pains to distance himself from those. Nowhere is this more apparent than in one of his last papers (1977*a*). In his own words:

> The Keynesian idea that capitalism typically produced mass unemployment . . . expressed itself in the belief that less developed countries were characterized by large-scale "disguised unemployment," an assumption that incidentally justified all sorts of import-substituting and protectionist policies by its implication that labor for domestic production had a zero social alternative-opportunity cost. Its more important implication, however, was that development planning could safely and should necessarily concentrate on the accumulation of capital, in the narrow sense of physical productive equipment, structures, and "infrastructures." In this connection, . . . the Harrod-Domar equation ($g = sk$, where g is the growth rate, s the savings ratio, and k the capital-output ratio) provided the essential skeleton of a model for development planning. The growth rate could be maximized by maximizing the marginal saving from output growth and minimizing, through choice of techniques and expanding sectors, the "incremental capital-output ratio."
>
> . . . The "Mahalanobis model," derived from Soviet planning and the Marxist model of deve
> the choice between investing in ma
> tion goods and in machines to mak
> sumption goods . . . and demonst
> "heavy" as against "light" indu
> [ultimately] result in a . . . faster
> tion. . . . The apparently startling
> largely due to the fact that the concentra
> between . . . "heavy" and "light" industry successfully concealed the fact that the real choice at issue was between a higher and a lower savings ratio, that is, between more consumption later and more consumption now.
>
> The inapplicability of the Keynesian assumption of the availability of masses of unemployed, appropriately skilled labor was gradually recognized as a lesson of hard experience. Its positive consequence was an emphasis on "manpower planning" and manpower-requirement forecasting as a supplement to savings-investment planning and forecasting—a vogue which later encountered increasing criticism

> on the same grounds as the savings-investment planning approach had already encountered, namely, the assumption of fixed coefficients between "capital" ("physical" in the earlier case, "human" in the later) and output. Its negative consequence was a continuing effort to provide theoretical support for the concept of "disguised unemployment" in the face of criticism of it on grounds of inconsistency with economic rationality and utility maximization. [Pp. 363–64]
>
> The Harrod-Domar model['s] . . . successor was the Chenery-Strout "two-gap" model for analyzing and projecting development assistance needs and investigating the question of if and when development aid would cease to be necessary. The model envisaged two "constraints" on development investment, domestic savings and foreign exchange expenditure needs, the tighter constraint being binding. The model, aside from its assumption of rigid coefficients to the neglect of substitution possibilities, is logically inconsistent since it confuses two different sources of foreign exchange—earnings, which absorb domestic resources in production, and gifts, which add resources to domestic savings. [P. 368]

The preceding direct quotations leave no doubt as to Harry Johnson's position with respect to the various mechanistic approaches to the analysis of economic growth. It is less easy to get an equally succinct statement, in his own words, of his own position. This is understandable, since his style, both of writing and of thinking about problems, tended to be interwoven like a fabric rather than linear and aimed at a single point.

Nonetheless, one can construct a fairly clear picture of where Johnson stood. His starting point was the neoclassical one-sector growth model with technical advance, about which he wrote a nice article (1966*a*) expounding the basic model and extending it to include the monetary as well as the real side. The essence of this model is that economic growth comes from additions to the stock of basic resources and from changes in technique. He emphasized that the role of physical capital accumulation had been grossly exaggerated in the literature and spoke favorably of the tendency for economic research to attempt to break down "the residual factor in economic growth" into components such as economies of scale, improvements in the quality of productive factors (through education in the case of labor and through improved efficiency in the case of machines), and shifts from a less efficient to a more efficient allocation of resources.

As his ideas evolved in this area, his vision of the process became richer and more detailed, with excursions into the relevance to the growth process of migration, of patents and other incentives to innovation, and of educational policy choices. The general tenor of his thinking on the process of growth can be gleaned from the following quotations taken from his 1963 lecture, "Economic Growth and Economic Policy" (1963*a*), which aimed at presenting "the main outlines of a policy aimed at promoting economic growth in a free enterprise economy":

> Such a policy would involve reform of the tax system to eliminate elements that inhibit growth and promote inefficiency. [Appropriate reforms] would entail replacing income by expenditure as the basis for personal taxation [so as to] eliminate the double taxation of saving; [integrating corporate and personal taxes with] corporate earnings being attributed to stock-holders and taxed as personal income; and finally [eliminating] the tariff, [using] subsidies . . . from the regular government budget if certain industries are deemed of special national importance. [Pp. 61–62]
>
> A second . . . series of measures would . . . increase the efficiency of the market system, particularly by improving the mobility of labour between regions and between occupations, [reforming] the educational system [and initiating] programmes for the retraining and relocation of labour [P. 62]
>
> Tax concessions for industrial
> [are not advisable because they] n
> tortions into the functioning of the
> stitute the judgments of the legisla
> the competitive system. . . . To t
> [evidence] has not . . . substantiated
> nal economies [whereby such investme
> benefits that do not accrue to the enterprises undertaking them. [P. 63]

III. Economic Policy in the Developing Countries

Harry Johnson's thoughts on economic policy in the LDCs can be approached either through his contributions to the general literature of economics or through those that focus specifically on the LDCs. His general contributions reveal him as virtually embodying the neoclassical tradition in economics; his more specific comments on the LDCs show him as an acute and astute observer, wise though at times sharp and acerbic in his judgments.

Good examples representing the neoclassical tradition are his related papers on "Tariffs and Economic Development" (1964) and "Optimal Trade Intervention in the Presence of Domestic Distortions" (1965*a*). Some of his key conclusions in these treatments are:

> The economic analysis of the various arguments for protection . . . can be summarized in two central principles: (1) Only the optimum tariff argument provides an economic justification for tariffs: all other arguments for protection are arguments for subsidies. (2) The use of the tariff on the basis of any of the other arguments may make matters worse rather than better. . . . whether . . . the tariff increases or decreases real income depends on . . . various [parameters] . . . and cannot be determined by *a priori* reasoning. [1964, p. 8]

> Welfare maximization requires a correction of [a] domestic distortion by an appropriate tax or subsidy on production, consumption, or factor use, and not a tax or subsidy on international trade; [moreover], given the presence of a domestic distortion, protection designed to offset it may decrease welfare rather than increase it. [1965*a*, p. 31]

> For the [infant industry] argument to be valid, . . . it must be demonstrated . . . that the social rate of return exceeds the private. . . . [One reason why this might occur stems from] the fact that . . . once knowledge of production technique is acquired, it can be applied by others than those who have assumed the cost of acquiring it. . . . The other reason why the social benefit may exceed private hinges on the . . . technique of production [being] embodied in the skill of the labor force [in which the workers have] property rights [even when they were] acquire[d] at the employer's expense. . . . [Whichever of these reasons might apply], it is apparent from the general principles governing optimal governmental intervention . . . that . . . some sort of subsidy to the infant industries, rather than protection [is indicated]. [1965*a*, pp. 27–29]

> The recommendation of a uniform tariff rate on commodities whose production it is desired to protect . . . has no theoretical validity. . . . For consistency, [a uniform rate rule]

> must comprise all imports. [When it does so, such a rule makes sense] . . . for a . . . policy aimed at increasing self-sufficiency, [for a] uniform tariff rate applied to all importable goods will . . . minimize the total excess cost of import-saving. [1964, pp. 18–19]

The preceding excerpts illustrate very well the flavor of that part of Johnson's theorizing that most directly bore on the policies of less developed countries. But they are unquestionably pieces of theory. There follow a few excerpts reflecting Johnson as an observer of and commentator on the policy behavior of the LDCs, taken from his (1967*b*) Brookings book, *Economic Policies toward Less Developed Countries.*

> Economic planning, implemented by controls of all kinds over the allocation of investment resources, commodities, and foreign exchange, is characteristic of less developed countries, . . . [and] in practice . . . give[s] rise to all sorts of inefficiencies and wastes of resources. . . . Licensing systems inevitably favor the established enterprise against the new competitor and hamper the ability of the more efficient firm to [expand]. . . . Fair distribution of scarce licenses may [require] producing far below capacity . . . ; discriminatory distribution . . . may protect the inefficient. . . . [But] the main impediment . . . imposed by licensing . . . is its displacement of the market mechanism [by an] administrative mechanism of detailed quantitative decisions[1] [P. 69]

> Development plans typically steer [illegible] of the available . . . resources towa[illegible] development policy . [illegible] depresses [illegible] [by raising] the prices of industri[illegible] [and by holding] down the prices rece[illegible] producers. . . . [Where] an export surplu[illegible] products [exists], it is generally deliberate policy to tax their producers heavily, [reducing] export earnings [and encouraging] the development of alternative supplies from elsewhere. [Pp. 70–71]

> [On industrialization via import substitution], it is particularly necessary to stress the importance of the input-output relations in a modern industrial structure and the effects of protection of inputs in raising the cost of production of the user industries, rendering them unable to compete in the

> world market. [The establishment of] a "new international division of labor" [requires] the abandonment of the effort [by LDCs] to achieve self-sufficiency. [Instead they should] favor . . . specialization on a few processes or products, allowing the producers to take full advantage of the availability of materials, components, and equipment in the world market.
> . . . The problems [above] of inefficiency and noncompetitiveness [are often exacerbated by] inflationary monetary and fiscal policies in combination with unwillingness to devalue the currency. [This unwillingness causes import restrictions to be] guided by political expediency and particularly by the priority accorded to investment goods and materials over consumer goods (especially "luxuries"). . . . the resulting protectionist system is extremely inefficient. [P. 74]

> [Apart from their influence on protectionism] inflationary fiscal and monetary policies have a variety of deleterious effects on economic development and efficiency. The most serious are, first, the distortions . . . caused by effort to protect certain segments of the population from the effects of inflation, for example by holding down the domestic price of food or of urban transport to shield the industrial worker, or by holding down interest rates to channel real income to manufacturing firms; and second, the disturbance of the normal processes of investment decision by extreme uncertainty about the short-run rate of inflation to be expected. These effects of inflation are harmful both to economic development . . . and [particularly] to . . . development [based] on industrial exporting. [Pp. 75–76]

> [Of course], endemic inflation . . . almost invariably [stems from] political inability to agree either on the taxation required . . . or . . . on the division of the national income among the claimants to it. . . . The insistence of outside advisers on conservative monetary and fiscal policies [often] amounts in reality to a demand for political stabilization and generally a demand for resolution of political conflict in favor of one of the disputants. [P. 76]

IV. Economic Policy in the Developed Countries

In this section we deal with how, in Johnson's view, the policies of the developed countries with respect to trade in goods and services impinge on the economies and in various senses the welfare of the

LDCs. The discussion is restricted to issues concerning trade in goods and services, because those concerning movements in factors and technology will be dealt with in the next section. The treatment of multinational "solutions" of various kinds (commodity agreements, preferences, etc.) is similarly deferred.

Within the purview of this section there are two main points: (*a*) the various mechanisms by which the policies of the industrial countries make the passage of the LDCs to higher levels of development more difficult; and (*b*) the costly extra burdens that the foreign assistance has typically placed on the recipient countries, or, put another way, the apparently endemic (yet unnecessary) inefficiencies that characterized the foreign aid process as Johnson observed it.

On tariff structures he had this to say:

> In the general case, failing concrete evidence to the contrary, it is reasonable to assume that the escalation of tariff rates with stage of production typical of modern tariff structures implies a corresponding but steeper escalation of effective rates of protection with stage of production. . . . So far as the pattern of world trade is concerned, [escalation induces a] bias [of] trade toward raw materials, fuels, and semi-fabricates; towards producers' goods (capital goods) rather than consumers' goods; towards goods of a luxury nature capable of bearing high tariffs; and towards . . . goods distinguished by technical superiority sufficient to overcome the competitive disadvantage imposed by tariffs [1965b, pp. 20–21]

> The escalated tariff structures of th
> viously create serious barriers to
> underdeveloped countries on the
> relatively cheap labour, and instead
> for development . . . towards depen
> of natural resources. [P. 21]

> The escalated tariff structures of the advanced countries are therefore a potentially powerful inhibitor of economic growth in the underdeveloped countries. This may be especially so with respect to industrial products that demand a mass market to permit efficient exploitation of economies of scale, a market not available within the underdeveloped countries themselves. [P. 23]

But if the escalated tariff structures bear against the healthy industrial development of the LDCs, the use by the industrial countries of various devices to protect their own agricultural sectors tends to drive down the world prices of similar products produced by the developing countries. Johnson cites in various places (1967*b*, p. 86) D. Gale Johnson's calculations of agricultural prices received by farmers in different advanced countries over the prices that would have had to be paid for the corresponding goods if imported. Harry Johnson's own work in this area consists of calculating the effects of reducing sugar protectionism of various types. In one of his very few excursions into substantive work with actual empirical data, Johnson calculates the estimated consumption cost of sugar protection (cum excise taxes) in a number of countries, together with the production cost of such protection for some of the leading industrial countries. He summarizes:

> In the major (Western) protectionist countries, substitution of imports for protected domestic production, allowing for the effects on sugar prices of world free trade in sugar, would have saved the expenditure of real resources worth $318.7 million [at 1959 prices]. The additional export earnings of the exporting countries due to this substitution would have been worth $675 million . . . and the estimated net benefit from this additional trade would have been worth in the neighbourhood of $120 million. Altogether, with respect to substitution in production alone, and entirely ignoring the effects of the expansion of consumption that would have ensued on the adoption of free trade in sugar, $438 million of additional resources could have been freed for development purposes from these seven countries alone through the abandonment of protection. . . . A part, and in some cases more than the whole, of this saving would be automatically transferred to the exporting [LDC] countries through the assumed increase in the prices of existing [sugar] imports. [1966*b*, pp. 39–41]

> [Without doubt] the prevalence of sugar protection has substantial effects both in wasting resources and in reducing the earnings of the less developed countries that have a comparative advantage in sugar production. [1966*b*, pp. 41–42]

On the matter of foreign aid, particularly bilateral aid, Johnson was extremely disturbed by the economic inefficiencies it entailed—the more so because he saw these inefficiencies as utterly unnecessary and unwarranted, as counterproductive from the standpoint of the ostensible purposes of foreign assistance, and as self-serving on the part of the donor countries.

> [A] serious set of problems arises from the interaction of a number of elements in the present system of supplying aid: the predominance of bilateral aid, the predominance of project over program aid, and the practice of tying aid to purchases in the donor country. . . . The chief sources of [the resulting] inefficiency . . . are as follows.
>
> First, less developed countries may be able to obtain aid for projects of a "display" or "monument" type because some country likes to finance that type of project, but not for other projects that have more . . . value because no donor is interested. . . .
>
> Second, less developed countries may have to pay substantially more than the competitive world market price for equipment [financed by aid]. Consequently an element of inefficiency (artificially high capital costs) . . . is built into the project. . . . Moreover . . . [because of these excess costs] repayment of the loan will involve returning more real value than was received, and this may far more than offset any elements of "softness" in the loan. . . .
>
> [Third, in addition to paying more for aid-financed goods], less developed countries may have to be content with less expert advisers than are available in the world market and, perhaps more important, accept a technology that is not well adapted to their economic conditions and especially their relative factor-scarcities. [1967*b*, pp. 81–82]

V. Factor Mobility and Economic Developm

When Johnson advised developed countri
matters, he would urge them to introduce m
to relax restrictions on immigration, and to
free movement of capital. The tighter were
laws and the stricter its controls on capita
liberal it should be toward accepting imports fr
tries and toward providing aid in the form of capita
felt that restrictions on factor movements lay at the core of many policy problems of international economic relations. His view was internationalist, humanitarian, and cosmopolitan. Opposed to it was the chauvinistic nationalism of the self-seeking elites of the nation states (large and small), into which he saw the world as having been accidently and arbitrarily split. He wrote extensively on three aspects of this debate: the brain drain—or labor mobility; the role of multinational corporations; and technology transfer

Labor Mobility

Johnson's position on labor mobility, or the "brain drain," as it manifested itself in the development debate in the 1960s, is reflected in the following quotations.

> My objection to the [concept of the gap between rich and poor countries] as the focus of analysis is not merely that the use of statistical averages can produce unnecessary paradoxes or nonsense results, but that it assumes the nation as the unit of analysis rather than the people that are born and live in a world divided into national states, and hence biases thinking towards a nationalist rather than a cosmopolitan view of the nature of the problem and solutions to it, and towards a neglect of the welfare of the individuals who better themselves by moving. No one seriously disputes the economic benefits, in the form of both improved allocative efficiency and competitive pressures for increasing productivity, that result from the mobility of labour within the nation; in fact nations typically seek to increase labour mobility by public policy. But the same type of movement of labour in search of better opportunities, but between rather than within nations, is widely regarded as a serious problem, damaging to the economic development of the poor nations.
>
> If one takes a cosmopolitan point of view, there can be no doubt that efficiency, growth and the relief of world poverty would be very effectively promoted by unlimited freedom of migration . . . The prevalence of poverty among the nations, and the need for transfers of capital and knowledge to help remedy it, can be attributed to an important extent to the barriers to immigration imposed by the advanced countries, coupled with tariff and other barriers to trade that restrict "indirect immigration" from poor to rich countries via the export of goods and services. [1972, pp. 381–82]

Replying to criticism of the "gap" paper, to the effect that being in favor of free migration meant that he supported an unequal distribution of world income, he said that, on the contrary, "in believing that emigration was a good thing, he was trying to improve income distribution. Migration laws were contrary to improving income distribution. They prevented people emigrating and were often accompanied by barriers to removing such differences by trade" (1972, p. 408). In the body of the paper (pp. 389–91) he dubbed as "bad economics" the argument that alleged market failures (with human capital not being paid its full social value) justified some form of restriction of the free emigration of educated citizens out of poor countries. He accepted that

> the ease of migration of educated people from underdeveloped countries, especially those in which English is the language of instruction, to advanced countries is a serious limitation on the potentialities of achieving economic development by educational investment and suggests the social desirability of devising means of obliging either the emigrants themselves or the countries receiving them to repay the social capital invested in them to their countries of origin. [1963*b*, p. 233, n. 6]

But he insisted that "compulsion should be confined to the subgroup of students . . . whose education is paid for by their own or an advanced-country government as a social investment in the promotion of economic development" (1972, p. 390).

The one qualification to his advocacy of freedom of migration arose from his Malthusian fear that the populations of some developing countries might breed to the subsistence level, quickly replacing the emigrants. Indeed, emigrant remittances (which on the whole he considered to be a much overlooked benefit of migration) could in an extreme Malthusian case produce a population multiplier effect. Since optimal family size is a function of expected income, which in turn is a function of education, he believed that developed countries could make "a constructive response to less-developed-country demands in the indirect forms of substantial support for population control policies and for programs of mass ele
an emphasis on vocational training" (1976*b*

Multinational Corporations

In the analysis of international economic pol
a great deal of importance on the increasing
as a political entity, the various forms of private
nomic entities, and the economic welfare of the individual (see his papers in *Economic Nationalism in Old and New States*, 1967*a*). He recognized that serious conflicts can arise "in the relations between the nation-state as a powerful political entity with a bordered geographical domain and the multinational corporation as a powerful economic entity with an unbordered world market domain" (1970, p. 28). But he felt that private enterprise in the form of the multinational corporation could, while seeking its own private profits, contribute significantly to the economies of the less developed countries. He thought that the planning approach to development had failed and that offi-

cial aid, designed to support this approach, would decline. The following quotations indicate how, in his view, multinational corporations could contribute to development:

> While the multinational firm has an incentive to invest in the transformation of the local economy only to the extent that such investment promises greater profits, the incentive may nevertheless be sufficient to induce a substantial contribution to development. Two particular incentives are especially important in this connection. First, labor skills imported from the developed countries are extremely expensive in comparison with the cost of training local labor, especially as the cost of training is largely either the labor-time of teachers or the labor-time of students, both of which are infinitely cheaper in poor countries than in rich ones. The foreign company will thus have a large profit incentive to train a local staff rather than import foreign labor.
>
> Second, given the complex input-output relations characteristic of modern industry and the dependence of profitable utilization of the product on the knowledge of the user, the firm may have a substantial incentive to invest in the diffusion of productive knowledge in two directions: to the local suppliers of the inputs needed in its production process and for which it demands quality standards superior to the customary standards of those industries and to local customers who have to be taught the technology of using the firm's products effectively. [1970, pp. 26–27]

> Any more general influence in promoting development will have to stem, on the one hand, from the exemplary value for local enterprises of the existence of efficient, well-managed, science-based subsidiary firms and their production of skilled local workers and executives who can be attracted into local enterprises and, on the other hand, from the ability of the state to use corporate and individual income taxes levied on foreign firms and their employees to finance education and other developmental expenditures. [1970, p. 27]

He believed that "the limitations on the potentialities of foreign direct investment as an agent of economic growth derive in part from social considerations that are generally accepted as just by well-intentioned people but in fact serve to inhibit the development process" (1970, p. 27). Pressure for wages and working conditions comparable to those in the home country of the company comes in this category. Other limitations come from a variety of other pressures: to use local inputs, to hire nationals beyond the point the company itself considers efficient, to meet export targets, and to conduct research and development locally. All these can misallocate the resources involved and raise production costs—sometimes to the point of dissuading potential investment or even provoking disinvestment.

Johnson considered that some of the problems encountered by multinational corporations represent LDC reactions to the practice (in some countries outside the United States) of influencing the activities of large businesses "covertly by conspiracy and backscratching among specialized members of a recognized and homogeneous political, economic and cultural national elite" (1970, p. 29), rather than through the law. He also recognized that competition among developing countries for the investments of multinational corporations could eat away at the potential development contribution that those investments could make.

Technology and Economic Development

Johnson's views on technology are easy to identify once his fundamental positions are understood:

a) "[Optimal output in the competitive model requires] that everyone has access one way or another to comparable technology—through competition in the supply of goods embodying such technology, or through migration of either capital [illegible] through the free availability of knowledge [illegible] 'oral tradition' " (1973*b*, p. 11).

b) "Technology is a form of capital, and [illegible] investment" (1977*c*, p. 322).

Against this background he argued that

> society is faced with a dilemma—th[illegible] knowledge costs resources, and investme[illegible] for somehow, but that optimization of its use requires that it be freely available to all potential users. The uneasy compromise society has chosen, at least for most commercially useful knowledge, is to grant a limited monopoly to the inventors of new knowledge, protected either by patents or by commercial secrecy, and trust to time and competition gradually to erode the monopoly. The result is both an alliance of technological superiority with the ownership of capital, especially in the large national and multinational corporation, which generates political and social tensions, particularly in less developed countries playing "host" to direct foreign investment in

> order to acquire better technology; the persistence of long technological lags among the more and the less advanced countries both reflects the slowing down of the diffusion of new technology by monopoly, and preserves inequalities of factor prices. One should observe, however, that such technological lags among nations may reflect the existence of barriers to the free movement of goods, capital and labor, rather than anything inherent in the nature of technology itself. If, for example, there were no barriers to the free movement of goods, one would expect production with a new technology to locate itself in regions with the lowest cost of factors of production; and if capital were free to move, these would be the regions with the lowest labor costs. But with barriers to trade on both sides, the transplantation of technology for production for export back to the home market is impeded; production for sale in the market of the "host" country becomes profitable only when factor prices, tariffs, or income levels make it profitable—and the transplant can easily be uneconomically premature. [1973*b*, pp. 41–45]

Johnson did not regard such restrictions on the development of new technology–based trade as necessarily involving a welfare loss for developing countries. For example, he argued that

> competition within and between . . . countries with surplus or elastic populations tends to drive prices down to subsistence, on classical unlimited-labour-supply lines; and this is likely to be the result of unlimited competitive expansion of exports, initially made highly profitable by the diffusion of technology to and among the developing countries.
>
> The fundamental solution to this problem obviously lies in the control of population growth in the less developed countries, pending ultimate arrival of the stage of development at which procreation is controlled by voluntary choice of the number of children the parents can afford to feed, house and educate to their own or better standards of productive contribution. In the meantime, control over the speed of development of trade based on technological diffusion may, intentionally or unintentionally, have the effect of contributing more to the raising of living standards and the strengthening of automatic economic pressures for development than cut-throat competition among developing countries in exporting technologically-advanced products to the developed countries. [1975, p. 159]

He did, however, feel that "where labour is prevented from moving to [an] advanced country by restrictive immigration laws, the transmission of technical knowledge can be regarded as a form of compensation for damages" (1961, p. 12, n. 21).

VI. Institutional Aspects of International Economic Relations

Johnson had strong feelings about the international economic relations issues covered in the North/South debate, feelings that he made no attempt to hide and that crop up frequently in his writings. His position rested on three convictions. First, he accepted that a moral case can be made for transferring resources from the developed to the developing countries. This case could be either based on humanitarian concern or justified as compensation for the injuries inflicted on the LDCs because past and present interventions in the market by developed countries had been and were biased against the trade and investment interests of developing countries. Second, he believed that many of the biased and injurious interventions resulted from the fact that policymaking in both developed and developing countries was often in the hands of elites that were at best nationalistically chauvinistic and at worst, but all too frequently, self-serving and self-preserving. The policies of such elites only rarely promoted the best interests of the citizens of developing countries on whose plight the rhetoric of development focuses attention. And third, as he demonstrated so frequently and so conclusively
proposed for the new international econ
ceived and analytically unsound. And many
from the secretariats of the international o
UNCTAD); these he saw as deriving largel
and carrying their predominant attitudes
self-serving international supra-elite. Thre
which Johnson felt particularly strongly were int
ity agreements, trade preferences, and the link between the creation of international liquidity and aid.

International Commodity Agreements

Johnson had strong views on international commodity agreements. He considered them to be "[one of the] ideas of Raúl Prebisch, especially as institutionalized and vulgarized through UNCTAD, [which] have become an increasingly powerful obstacle to co-operation in the promotion of the development of the developing countries, in the specific sense that sympathy with aspirations for development has to be demonstrated by the acceptance of economic nonsense and the endorsement of proposals that . . . maximize the prospective costs and minimize the prospective return" (1976*b*, pp. 329–30). He believed that economic theory and experience teach us that commodity agreements "aside from the difficulty of devising and operating them—are an exceedingly doubtful instrument for promoting economic development" (1976*b*, p. 327) and that

> economically, such agreements . . . amount to an extremely cumbersome method of transferring income from consumers to producers, a method that entails considerable economic waste through the distortion of consumption and production decisions by arbitrary high prices, and through the probable necessity of production restrictions and surplus disposal. In addition . . . there are not many commodities for which the long-run demand is sufficiently inelastic for price-increasing measures to increase total revenue; and . . . price-fixing is a very inefficient method of providing equitable incomes for agricultural producers. It would be economically far more rational to provide income transfers directly from the rich to the poor countries, rather than to seek the same effect through the cumbersome and probably ineffective route of international commodity agreements.

Although he allowed that in a few cases increased transfers could under very special circumstances be effected by some forms of agreement, by the mid-seventies he was increasingly disillusioned with attempts to "disguise" transfers. He had come to the conclusion that "the successor governments of the independent developing countries have . . . become much more sophisticated, in the sense of appreciating that there is no reason why high prices for commodity exports should be passed along to the stupid farmers, instead of being creamed off for spending on and by the educated government bureaucracy in the name of promoting economic development, social justice, or what have you" (1976*b*, p. 320). Apart from this, he believed that the continued popularity of international commodity agreements is a matter of faith which "as is usually the case, . . . rests either on ignorance of past history or the obstinate belief that what went wrong last time was attributable either to lack of will or cleverness, or unwillingness to commit sufficient financial resources to the enterprise—but never to inherent difficulties that could be understood in terms of elementary economic analysis" (1976*b*, p. 322).

Trade Preferences

Johnson's initial position on preferences was fairly straightforward. He felt that "the clear implication of the theoretical considerations and empirical evidence is that, despite impressions to the contrary derived from contemplation of nominal tariff levels, preferences in developed-country markets for exports of manufactures and semi-manufactures might well exercise a powerful influence in expanding the export earnings and promoting the industrialization of the less developed countries" (1967*b*, p. 173). This position was based partly on the assumption that the "formation and policies of the European Common Market will block further progress after the Kennedy Round toward trade liberalization along traditional GATT lines" (1967*b*, p. 205) and partly on the assumption that any preference scheme negotiated would be fairly liberal in terms of country and product coverage and in the extent of the preferential margin extended.

Johnson was wrong on both counts. In fact, the Europeans did not stop the process of liberalization in GATT. Moreover, the preference schemes that did emerge were ill designed and were regarded by Johnson as at best third- or fourth-best policy instruments, or at worst—in the form adopted by the EEC for the African/Caribbean/Pacific Group of States (ACP)—as a calculated move by the Europeans designed to keep their ex-colonies in a "neo-neo-colonialist" dependent relationship. He felt that the developed countries should not be encouraged to bend the rules of international trade in order to establish such schemes.

> The less developed countries would be b
> sist, so far as they can, that the d
> accept and live by the principle
> than to tolerate rich country vio
> attempt to counteract these viola
> own, coupled with demands for
> foreign aid and trade preferenc
> should be] to get the rules change
> ing a closer approximation to honesty
> developed countries should have a strong concern with the promotion of freer world trade on a global basis, and especially with respect to agricultural products, and should not allow themselves to be diverted into pursuit of such gimmicks as special trade preferences [1973*a*, p. 34]

In addition, he became more and more convinced that his original position, that the potential benefits of trade preferences might not be realized because of self-imposed supply constraints, was correct and in fact the case. In the Brookings volume he had said:

> It is necessary to determine what factors account for the inability of less developed countries, and specifically of the

> "developing" countries that already produce manufactures for the home market, to export in competition with the developed countries in spite of their comparative advantages . . . of materials and low-wage labor, and how significant these factors are empirically. . . . [In Johnson's view, the culprits were] the import-substitution and currency-overvaluation policies typically pursued by the governments of less developed countries. . . . [To his mind the cost disadvantages resulting from such policies were] frequently . . . far greater than the competitive advantage that could be conferred by preferences from the developed countries [implying that] neither preferences nor nondiscriminatory tariff reduction would help the less developed countries unless they were prepared to make major changes in their tariff and exchange-rate policies. [1967*b*, p. 206]

The "Link"

The various proposals to link the growth of some new form of international liquidity to the provision of aid to developing countries elicited a strong response from Johnson. He recognized that the developing countries could benefit from a reform of the international monetary system but did not feel that the various proposals for a link between the growth of liquidity and aid were a sensible move. He criticized them initially on the grounds that "there is no obvious advantage, and much evident and avoidable complexity, in attempting to solve two different and incommensurable problems with one and the same institutional change" (1967*b*, p. 223). He argued that most of the proposals were simply hoping surreptitiously to increase the total flow of aid to developing countries and did not think them likely to succeed. "[This] assumes that the aid policies of the developed countries are independent of international monetary operations, so that whatever aid would be channelled by a world central bank to the less developed countries would be a net addition to their aid receipts. Since in the real world the left hand usually knows full well what the right hand is doing (though it may not admit it publicly), this is an unrealistic assumption" (1967*b*, p. 228).

Johnson's second criticism—that the proposals saddled one instrument with two potentially conflicting objectives—was two pronged. On the one hand, the amount of new international liquidity required in any period by the international economy might have no relation to the requirements of developing countries for aid in that period. On the other hand, there was no way of ensuring that a distribution of funds to less developed countries according to their estimated liquidity needs would correspond either with their capacity to absorb capital in economically sound projects or with the objectives of the donor countries. Making the aid objective paramount could result (through excessive global monetary expansion) in an unstable international economy. This, of course, would conflict with the long-run interests of the developing as well as the developed countries. He felt that the international monetary system as it was in the mid-1960s fragmented the international economy by pushing developed countries that faced (or were frightened at the prospect of) balance-of-payments problems into adopting policy measures that restricted the inflow of imports and the outflow of capital. He regarded both of these as being inimical to the interests of developing countries. He argued that

> the less developed countries would benefit . . . from the establishment of an international monetary system that provided a more adequate combination of liquidity and adjustment mechanisms. The most important contributions of such a system would be to remove the inhibitions to expansion of aid generated by fear of consequent balance-of-payments difficulties and to facilitate the untying of aid, thus increasing its efficiency. In addition, less developed countries would benefit significantly from the removal of restrictions dictated by balance-of-payments difficulties, whose effect is frequently to protect production in the developed countries and to aggravate un
> trade.
>
> The less developed countri
> important interest in the estab
> monetary system that would p
> production and trade at high l
> tivity by providing an adequate s
> tional liquidity [1967*b*, p. 215]

In the Brookings book he went farther than this, asserting that the developing countries "have a particular interest in the establishment of a system that will expand international liquidity at a rate great enough to impart an inflationary bias to world economic development" (1967*b*, pp. 216–17). Ten years later, however, he was more worried by inflation. He felt that the "favorable climate for economic development [in which] . . . the developing countries [have] the advantage of trading with an advanced-country world characterized by fairly steady sustained economic growth at fairly stable (more accurately, slowly rising) prices . . . has been destroyed, at least temporarily, by world inflation" (1976*b*, p. 321). This world inflation was to some extent the result of the International Monetary Fund's effectively introducing a disguised and limited form of "link" as a result of its attempts to

> strengthen its political support and acceptability by catering to the special interests of politically powerful groups of constituents. Specifically, the Fund has been cultivating support . . . by potentially extremely inflationary devices. . . . It has been steadily enlarging its role as a lender of international liquidity on concessionary terms to developing countries in balance-of-payments difficulties. . . . The danger is that the developing-country group will be so fascinated with . . . the age-old fallacy that the creation of money is a way of getting something for nothing—something real for something paper—that they will provide indirect support for international . . . monetary policies that [perpetuate] world inflation. [1976*b*, pp. 331–35]

VII. The Human Side

Harry Johnson's introduction to the world of developing countries came in 1956, when he was invited by the International Economic Association to teach refresher courses in Karachi and Singapore. This opened the door to a series of visits to Asian countries and to a great many abiding personal connections with Asian economists. (He later traveled to Ghana and made several trips to Latin America, lecturing in Argentina, Chile, Mexico, Panama, and Uruguay; nonetheless his deepest and firmest roots in the developing world were and remained in the Asian countries.)

Over the years he inspired many Ph.D.s from the developing nations, but he also had enormous influence on many who never studied directly under him. Few if any in our profession even came close to Harry Johnson in terms of willingness to read and comment on the papers of younger economists. This he did simply in his several roles as journal editor, but even beyond what he did "officially," he conscientiously read and commented on the work handed to him by younger economists at just about every institution he visited. A great many beneficiaries of his generosity in this respect were members of the post–World War II generation of professional economists from the LDCs.

Part of the human side of Johnson's links with the developing world lay in his selfless dedication of time, energy, and ideas to individuals and institutions there. Another part lay in the profound moral messages that permeate his writings concerning the developing countries. In his observations he was never naive, always realistic, never pusillanimous, always forthright. No one who reads his writings can fail to sense that here was a man who cared deeply about improving the physical well-being and the human dignity of the millions upon millions of individuals born into poverty and misery. Like other seers throughout history, he recognized the fallibility and corruptibility of men and institutions. His deep understanding of the human condition and his willingness to look unpalatable truths squarely in the face helped to give his judgments a special force and power. He was no simple preacher, but a seasoned, sage, and sound observer of the real world who saw, through many and inevitable obstacles, a path to bettering the lot of humankind.

His views on foreign aid reveal the interplay of realism and moral tone that characterizes much of his work. On the one hand he felt that foreign aid was a good thing because it increases the availability of real resources to societies with many needy people. On the other, it goes from government to government through fiscal systems that are often corrupt; sometimes it even ends up going from the poor (in donor countries) to the rich (in the receiving societies). The resulting waste, together with the stridency of demands by some representatives of the developing countries, leads to disillusionment within the donor societies and to a decline in the flow of aid. Therefore the "useful life" of foreign aid is limited—one must use it well to induce developing countries to reform their economic policies so that they will be able to take fuller advantage of the opportunities for long-run development that can follow from an intelligent application of economic principles.

Just as he did not mince words in talking a
forthright on the interactions of economic
This was partly reflected by his role as
Committee on New Nations at the Univer
ings it came through in many places, perha
and characteristically than in the followin
economic development is an integral par
development and social development consists of
increased personal freedom, which means destroying the powers of the *élites* and not merely switching the party or social class label of the *élite* that commands the power [of] the State" (1974, p. 9).

References

Johnson, Harry G. "The De-stabilising Effect of International Commodity Agreements on the Prices of Primary Products." *Econ. J.* 60 (September 1950): 626–29.

———. "Notes on Economic Development and the Maximum Rate of Growth." *Malayan Econ. Rev.* 2 (April 1957): 16–22.

———. "Effects of Changes in Comparative Costs as Influenced by Technical Change." *Malayan Econ. Rev.* 6 (October 1961): 1–13.

———. "Economic Growth and Economic Policy." In *The Canadian Quandary: Economic Problems and Policies.* Toronto: McGraw-Hill, 1963. (*a*)

———. "Towards a Generalized Capital Accumulation Approach to Economic Development." In *The Canadian Quandary: Economic Problems and Policies.* Toronto: McGraw-Hill, 1963. (*b*)

———. "Tariffs and Economic Development: Some Theoretical Issues." *J. Development Studies* 1 (October 1964): 3–30.

———. "Optimal Trade Intervention in the Presence of Domestic Distortion." In *Trade, Growth, and the Balance of Payments: Essays in Honor of Gottfried Haberler*, by Robert E. Baldwin et al. Amsterdam: North-Holland, 1965. (*a*)

———. "The Theory of Tariff Structure, with Special Reference to World Trade and Development." In *Trade and Development* (with Peter B. Kenen). Geneva: Libraire Droz, 1965. (*b*)

———. *Trade and Development* (with Peter B. Kenen). Geneva: Librairie Droz, 1965. (*c*)

———. "The Neo-Classical One-Sector Growth Model: A Geometrical Exposition and Extension to a Monetary Economy." *Economica* 33 (August 1966): 265–87. (*a*)

———. "Sugar Protectionism and the Export Earning of Less Developed Countries: Variations on a Theme by R. H. Snape." *Economica* 33 (February 1966): 34–42. (*b*)

———. *Economic Nationalism in Old and New States.* Chicago: Univ. Chicago Press, 1967. (*a*)

———. *Economic Policies toward Less Developed Countries.* Washington: Brookings Inst., 1967. (*b*)

———. *Trade and Aid Policies: The UNCTAD Alternatives.* Washington: Brookings Inst., 1967. (*c*)

———. *Comparative Cost and Commercial Policy Theory for a Developing World Economy.* Stockholm: Almqvist & Wiksell, 1968.

———. "Thrust and Response: The Multinational Corporation as a Development Agent." *Columbia J. World Bus.* 5 (May/June 1970): 25–30.

———. "Labour Mobility and the Brain Drain." In *The Gap between Rich and Poor Nations*, edited by Gustav Ranis. London: Macmillan, 1972.

———. "International Trade and Economic Development." In *Trade Strategies for Development*, edited by Paul Streeten. London: Macmillan, 1973. (*a*)

———. "Trade, Investment and Labor, and the Changing International Division of Production." In *Prospects for Partnership. Industrialization and Trade Policies in the 1970's*, edited by Helen Hughes. Baltimore: Johns Hopkins Univ. Press (for World Bank), 1973. (*b*)

———. "The Western Model of Economic Development." *World Development* 2 (February 1974): 9–10.

———. *Technology and Economic Interdependence.* London: Macmillan (for Trade Policy Res. Centre), 1975.

———. *Trade Negotiations and the New International Monetary System.* Geneva: Graduate Inst. Internat. Studies, 1976. (*a*)

———. "World Inflation, the Developing Countries, and 'An Integrated Programme for Commodities.'" *Banca Naz. del Lavoro Q.*, no. 119 (December 1976), pp. 309–35. (*b*)

———. "Changing Views on Trade and Development: Some Reflections." In

Essays on Economic Development and Cultural Change in Honor of Bert F. Hoselitz, edited by Manning Nash. *Econ. Development and Cultural Change* 25 (suppl.; 1977): S363–S375. (*a*)

———. *Money, Balance-of-Payments Theory, and the International Monetary Problem*. Essays in International Finance no. 124. Princeton, N.J.: Princeton Univ. Press, 1977. (*b*)

———. "Technology, Technical Progress and the International Allocation of Economic Activity." In *The International Allocation of Economic Activity. Proceedings of a Nobel Symposium*, edited by Bertil Ohlin, Per-Ove Hesselborn, and Per Magnus Wijkman. London: Macmillan, 1977. (*c*)

Harry Johnson as a Social Scientist

Richard E. Caves

Harvard University

"One of the penalties—or privileges, as the case may be—of advancing age and professional maturation in an academic career in the social sciences is that one is forced to think in terms of progressively broadening frames of institutional and cultural reference." So Harry Johnson (1975*b*, p. ix) prefaced his principal collection of essays on the social frontiers of economics. The widening sweep of his intellectual searchlight illuminated its targets not so much with originality of insight or breadth of social and cultural perception as with the penetrating power of his critical and analytical intellect. Johnson thought that the world revolved not from love but from individual utility maximization. These writings gain their unifying distinction from the remorseless application of its logic to interpreting economic and social behavior and criticizing economic policies. Although Johnson did not accord to the competitive market unlimited power to maximize human welfare, he did heap scorn on policies that took less than the full measure of rational behavior by the affected individuals and groups.

Another feature distinguishing many of these writings is their grasp of a large system of intersecting social and political forces. One reads with some surprise of Johnson's early fascination with Marxian analysis (Reuber and Scott 1977, p. 671). The Marxian postulates and analytical concepts were discarded long ago, but there survived a disposition to place social and political institutions on their foundations of the technology of production (and consumption, Johnson added) and the associated distribution of ownership claims and, in particular, to view political behavior as the outcome of economic interest groups.[1]

---

[1] Johnson (1974, p. 5) branded "clearly inadequate" the Marxian view that "the social, political, and ideological superstructure is determined by the technological and eco-

[*Journal of Political Economy*, 1984, vol. 92, no. 4]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9204-0005$01.50

This essay attempts to draw together Johnson's ideas on three subjects: changing social phenomena and their foundations in rational individual behavior, the operation of political systems and their effect on the substance of economic policy, and the working of the university and the intellectual marketplace.

I. Social Phenomena and Rational Individual Behavior

Many of Johnson's thoughts about the economic characteristics of modern society grew from his reflections on John Kenneth Galbraith's *The Affluent Society* (1958). Although Johnson felt no sympathy for Galbraith's central conclusions, he was intrigued by the book's central question: What are the economically distinctive features of a society that has reached unprecedented high levels of income per capita? It offers to individuals greater personal freedom and broader opportunities for self-fulfillment, interesting careers, and more satisfying rewards for the development of personal talents For most people it changes the basis for participation in production, so that it comes to rest on "personal knowledge, training, and education rather than on property ownership or the supply of labor time." The ascending importance of managerial ability and human capital in production processes makes every man a capitalist and contributes to social stability. The rising value of labor time also increases the capital intensity of household consumption Even military technology had responded to the rising economic and social value of human life, with the United States leading the effort to substitute th
rial inputs for human lives as the cost of
1975*b*, pp. 4–5).

Against these gains, Johnson felt that th
series of hazards on its members. The hous
of its important choices are investment d
periods, which demand complex calculat
formation. These investments are often hig
bad forecast can produce results much worse than
expected. Smaller family units and looser ties mean less self-insurance within the family. As a result, noncapable decision makers are in most risk-exposed positions. Long periods of education (investment in human capital) reduce the individual's sense of control over his life Affluence promotes and to some degree requires economic growth

---

nomic substructure, and specifically by the distribution of property rights associated with the system of production," because the "allegedly determinate superstructures react back on and change the economic substructure," and the technical substructure itself changes in response to human efforts to change and improve it and agglomeration; these bring a welter of externalities, with their potential for social conflict and market failure. These perceived social problems of affluence, Johnson felt, could be much clarified by application of such modern economic concepts as human capital, externality, and permanent income (1975*b*, chaps. 1–4).[2]

Human Capital

Johnson felt that the concept of human capital had "tremendous integrative power" to unify our understanding both of the labor market and of the household's decision problems. In the workplace the classical dichotomy between produced and original factors of production becomes increasingly meaningless. All factor inputs are best regarded as capital, their rewards consisting of a wage (interest and depreciation) plus a rent associated with the factor's ultimate scarcity (1975*b*, p. 43). The restrictive practices of professional organizations and trade unions are investments in the protection of quasi rents,[3] and "even the commitment of labor to the nationalization of certain industries can be interpreted as reflecting the desire to bring additional coercive power over the rest of society to bear in defense of the return to investment in . . . human capital" (1975*b*, pp. 20–21).

Johnson's concern with the acquisition of human capital dealt more sympathetically with the uncertainty and vulnerability of the investment than with the problems of income distribution that it poses. The ideology of a liberal democratic society supposes that the individual can be adequately equipped to cope, once provided with minimum education and socialized pensions. However, the complex knowledge required for prudent human-capital investments and the risks imposed by dynamic change call this into question. Personal risks are insurable in principle, but economic risks are unpredictable and hence uninsurable. The small size of the family unit condemns it to an undiversified portfolio of human capital, so that it is exposed to risks of illness, accident, and death, economic risks, and risks associated with inept policy choices. The family needs information and geographic and occupational mobility to make the most of its opportunities, but these are costly. Human capital in hand has some collateral value (consumer credit), but Johnson (1975*b*, pp. 47–49) felt that the inability to borrow on binding contracts against future quasi rents to human capital condemns this market to underinvestment in the absence of public intervention. Family pride in education and the opportunity to make intergenerational transfers without inheritance tax, however, cut the other way.

---

[2] The earliest of these essays appeared in 1960. Papers discussed in this essay often made their first appearance in hard-to-find sources; citations will therefore be made to convenient reprinted versions.

[3] Krueger (1974) was later to develop this notion.

Human-capital investments (on-the-job training excepted) cannot be financed like other investments, and Johnson wrote about various difficulties with the second-best procedures used instead. The provision of "free" university education has been undertaken by public authorities who often lose sight of the substantial remaining cost to the student of forgone income, and Johnson (1968) blamed some of the troubles of British universities in the 1960s on dissonant views of students and politicians about how much gratitude was due the latter from the former. An issue of more analytical substance arose in the "brain drain." Most countries underprice education and implicitly recoup the cost through progressive income taxes. Human capital is internationally mobile, however, and so it pays the individual once educated to migrate to a country offering higher pay or merely imposing a less progressive income tax. The country of emigration can lose because it then fails to collect tax revenue from the educated person equal to the cost of education supplied. An intergenerational transfer occurs if the current working generation pays to educate the next generation in the unfulfilled hope of being supported in retirement. A loss to society at large occurs, however, only if the incentive to migrate rests on differences in tax systems or if any externalities of educated people are greater in their birthplace than in the land to which they migrate (Johnson 1967*b*; 1975*b*, pp. 26–27).[1] Johnson felt that the national costs associated with the brain drain were generally overstated but that nonetheless they argue for load[illegible] cost of education on the student.

Income Distribution and Poverty

Johnson (1975*b*, chap. 14) regarded a pre[illegible] income distribution as an ethical choice b[illegible] analysis and the definition of poverty as [illegible] Nonetheless, he felt that economic reasoning could [illegible] the assessment of and prescription for each problem. To evaluate the maldistribution of income sensibly, one must apply the concept of permanent income. Different choices about early investments in human capital imply large differences in individuals' lifetime income profiles and, hence, some of the appearance of poverty in the distri-

---

[1] Johnson noted that the tendency to international equalization of the wages of educated people implied a widening of their scarcity rents in most countries (especially less-developed), with ensuing economic and social conflict.

bution of current income. Consumption needs also vary over the life cycle, to the same effect. The cross-section distribution of income reflects households' differing voluntary choices about their rates of saving and family formation, preferences regarding the balance of work and leisure and the regularity of the work pattern, and attitudes toward the riskiness of various possible jobs. One source of poverty in the affluent society is bad luck in the choice of specialized investments by individuals in skills or geographic location, a clear case for insurance through social assistance transfers (1975*b*, pp. 29–30).

The definition of poverty and policies toward it similarly need the application of economic analysis, Johnson (1975*b*, chap. 15) believed. Poverty is normally defined not by an absolute standard but by a moving one that maintains a fairly stable relation to average income. Economic growth tends to eliminate poverty only if it raises the share of productive assets owned by groups with initially lower incomes—not a likely outcome. The incidence of poverty also reflects utility-maximizing choices. Members of a multigenerational family group choosing to live under one roof might enjoy incomes on average above the poverty line, whereas some statistical poverty would appear if each generation maintains its own dwelling (1975*b*, pp. 28–29). Johnson was more concerned about poverty due to cultural restrictions on rational family planning and "weaknesses in the legal system governing the economic claims of mothers on the fathers of their children." If we are to make comparisons of current income, Johnson (1975*a*, pp. 12–15) warns that complex adjustments are necessary: labor income becomes current wages minus both interest on past investments in human capital and the costs of supporting the worker in retirement and raising children as replacements. Similarly, comparisons of the distribution of wealth ought to go beyond easily marketable assets to include, for example, the capital value of rights to the acquisition and maintenance of human capital provided free by the public sector. Poverty also results from discrimination, and Johnson (1975*b*, pp. 227–28) gave particular emphasis to discrimination against women. "In advanced industrial society, where reasonably good health and trained intelligence rather than muscular strength constitute the foundation of productive contribution, the traditional assignment of women to household management and child rearing is an anachronism." He held that society should choose between making women really equal, redesigning the family to relieve the differential burden it places on women, and acknowledging that the role of the household manager in maintaining the household is as exacting as the "breadwinner's."

The economist's preference for relieving poverty by outright income transfers clashes with the humanitarian's fondness for giving benefits in kind, and Johnson (1975*b*, pp. 232–35) sought to explain how the humanitarians come by their choice. The familiar paternalistic arguments assert that the poor should be induced to consume goods and services of which their humanitarian donors approve, so that "under the guise of improving the poor they are really seeking to create a society more deferential to people like themselves." Redistribution in kind thus serves to combat a negative externality for the middle-class majority from witnessing the behavior of people with preference functions different from their own. Better arguments for redistribution in kind, Johnson noted, are that they can partially correct for other market failures (food stamps as an offset to agricultural price supports, education countering imperfect markets in human capital). Also, they may help to avert the transference of poverty-prone attitudes from parents to children and the equipping of children with insufficient productive capacities.

Johnson (1975*b*, pp. 233–34) also noted that many poverty programs preferred by noneconomists presume that the poverty results from some social disequilibrium, so that those fallen into poverty could be restored with a relatively small social expenditure. Different conclusions tend to flow from economists' view of poverty as an equilibrium of choice and opportunity based on myopia, leisure preference, aversion to industrial discipline, or limited talents and capabilities.

Clear thinking about unemployment, he held, also requires recognition of the elements of individual choice lying behind it. The socially optimal value of any single index of unem
not just on labor-market frictions and many
on elements of voluntary choice. The lei
fected by unemployment benefits and social
factors as increased education, which eases
unemployment. Differences in leisure pref
higher long-run levels of unemployment
sure-labor preference itself is changed by the
affluent society. Technological change has predominantly raised the preference for leisure by lowering the relative cost of equipment used in leisure-time activities, but Johnson (1975*b*, pp. 22–23) was also intrigued by the possibility that some technological innovations in consumption might substitute material inputs for time in the creation of utility from leisure, thus raising labor supply. He was unimpressed by one economic influence asserted to affect the work-leisure choice—that of social security and the taxes to finance it (1975*b*, pp. 57–58); educated people are driven by the challenge of accomplishment, not the fear of hunger, and "work best when they have the security to concentrate on the job they are qualified to do." Also the

issue is a less urgent one in the affluent society with its revealed preference for increased leisure-time activities.

Externality and Property Rights

Johnson (1975*b*, pp. 5, 13) counted an increase in the perceived incidence of external diseconomies (pollution, congestion) as a market failure of the affluent society, because affluence is closely allied with overall economic growth and urbanization. But he relied heavily on the improved definition of property rights as the solution, with government interventions appropriate only where transaction costs frustrated a market solution. He heaped scorn (1975*b*, pp. 332–36) on people who would renounce present-day affluence because it allows too many others to compete with them for the private enjoyment of what we treat as public goods, yet he recognized (p. 13) a growing tension with the "common democratic presumption that every citizen regardless of income has a right to share in the original beauties of the natural environment, and that it is one of the functions of government to maintain and preserve that right." He emphasized once again the unifying role of capital theory for dealing with environmental problems. The environment provides a set of capital assets that can be either augmented or impaired by social progress, and their optimal exploitation or restoration can be managed only through investment decisions that properly value present and future benefits and costs.

II. Behavior of Political Decision Systems

In his technical writings, Johnson (1965, p. 256) expressed skepticism rather early about the traditional liberal posture of assuming that an identifiable way exists to maximize real income along social welfare lines and that the government wants to follow this course but needs only the analytical talents of an economist to tell it how to proceed. He thereafter showed recurrent interest in the behavioral approach to governmental decisions built with economic tools by Downs (1957) and others. Elected governments or political parties maximize votes in pursuit of various tangible rewards. The efficient development of their policy stances is hampered, however, by high costs of information, and so much information on voter preferences is injected by interest groups and lobbyists (and also the communication media). Likewise, the voters' investments in information about the policy stances of political parties are sharply limited, because the individual ballot has negligible power to influence political outcomes. Low-cost communication is provided by ideology, which reduces the detail that a party needs to relay to its potential adherents. Johnson (1967*a*, chap. 1) stressed ideology's role in shaping the political systems of less developed countries. Where voters' preferences for ideology are unimodal, as seems common in industrial countries, a two-party system tends to evolve with a good deal of imitation between the parties and jostling for the central position. Where democracy is less well established, however, the party in power has a strong incentive to forge a comprehensive and preclusive ideology in order to retain exclusive control of the government. The party's incentive for this strategy is greater in an LDC because the alternative job opportunities for office holders tend to be less good than in a developed country, where a larger cadre of administrative jobs can be found in the private sector.

Johnson used these ideas to explain many patterns of public policy, but perhaps nowhere more successfully than with the economic policy of nationalism. He encountered nationalist ideologies frequently in both LDCs and his native Canada and saw behind many economic policies the intersecting forces of nationalist ideology and interest-group pressures. Nationalism in the LDCs serves as an integrative force and the chief propellant of collective efforts to accelerate economic development, but Johnson (1967*a*, chap. 1) found it to exact serious costs by imposing policies that are probably inappropriate. One conspicuous component of LDCs' ideologies (and not just theirs) is a preference for industrial production over other forms of economic activity, which can express itself in policies to expand the industrial sector and depress the marginal return to the resources it employs below their best alternative uses. He showed formally (1965) how a preference for industrial output could affect the level of tariff protection chosen by a country, its particip
ing to reciprocal tariff reductions, and its
customs union. Reviewing actual patterns
suggested that specific industrial preferenc
what with a country's level of developme
both its comparative advantage and its as
seeking steel industries, more developed ones
and rather advanced ones the high-technology industries. His economic explanation of output preferences and related policies (1967*a*, pp. 6–9) drew upon the interest-group mechanism. Producer groups are the winners when the nationalistic ideology fosters a preference for placing status-giving jobs or property in the hands of nationals (see Breton 1964). This goal can be sought either by shifting the composition of output to create the preferred jobs or property (e.g., the preference for industrial output) or by removing these things from the hands of foreigners (restricting employment of foreign nationals or restricting the ownership of subsidiaries by foreign com-

issue is a less urgent one in the affluent society with its revealed preference for increased leisure-time activities.

Externality and Property Rights

Johnson (1975*b*, pp. 5, 13) counted an increase in the perceived incidence of external diseconomies (pollution, congestion) as a market failure of the affluent society, because affluence is closely allied with overall economic growth and urbanization. But he relied heavily on the improved definition of property rights as the solution, with government interventions appropriate only where transaction costs frustrated a market solution. He heaped scorn (1975*b*, pp. 332–36) on people who would renounce present-day affluence because it allows too many others to compete with them for the private enjoyment of what we treat as public goods, yet he recognized (p. 13) a growing tension with the "common democratic presumption that every citizen regardless of income has a right to share in the original beauties of the natural environment, and that it is one of the functions of government to maintain and preserve that right." He emphasized once again the unifying role of capital theory for dealing with environmental problems. The environment provides a set of capital assets that can be either augmented or impaired by social progress, and their optimal exploitation or restoration can be managed only through investment decisions that properly value present and future benefits and costs.

II. Behavior of Political Decision Systems

In his technical writings, Johnson (1965, p. 256) expressed skepticism rather early about the traditional liberal posture of assuming that an identifiable way exists to maximize real income along social welfare lines and that the government wants to follow this course but needs only the analytical talents of an economist to tell it how to proceed. He thereafter showed recurrent interest in the behavioral approach to governmental decisions built with economic tools by Downs (1957) and others. Elected governments or political parties maximize votes in pursuit of various tangible rewards. The efficient development of their policy stances is hampered, however, by high costs of information, and so much information on voter preferences is injected by interest groups and lobbyists (and also the communication media). Likewise, the voters' investments in information about the policy stances of political parties are sharply limited, because the individual ballot has negligible power to influence political outcomes. Low-cost communication is provided by ideology, which reduces the detail that a party needs to relay to its potential adherents. Johnson (1967*a*, chap. 1) stressed ideology's role in shaping the political systems of less developed countries. Where voters' preferences for ideology are unimodal, as seems common in industrial countries, a two-party system tends to evolve with a good deal of imitation between the parties and jostling for the central position. Where democracy is less well established, however, the party in power has a strong incentive to forge a comprehensive and preclusive ideology in order to retain exclusive control of the government. The party's incentive for this strategy is greater in an LDC because the alternative job opportunities for office holders tend to be less good than in a developed country, where a larger cadre of administrative jobs can be found in the private sector.

Johnson used these ideas to explain many patterns of public policy, but perhaps nowhere more successfully than with the economic policy of nationalism. He encountered nationalist ideologies frequently in both LDCs and his native Canada and saw behind many economic policies the intersecting forces of nationalist ideology and interest-group pressures. Nationalism in the LDCs serves as an integrative force and the chief propellant of collective efforts to accelerate economic development, but Johnson (1967*a*, chap. 1) found it to exact serious costs by imposing policies that are probably inappropriate. One conspicuous component of LDCs' ideologies (and not just theirs) is a preference for industrial production over other forms of economic activity, which can express itself in policies to expand the industrial sector and depress the marginal return to the resources it employs below their best alternative uses. He showed formally (1965) how a preference for industrial output could [illegible] protection chosen by a country, its partici[illegible] ing to reciprocal tariff reductions, and [illegible] customs union. Reviewing actual patterns [illegible] suggested that specific industrial preferen[illegible] what with a country's level of developme[illegible] both its comparative advantage and its aspi[illegible] seeking steel industries, more developed ones ar[illegible] and rather advanced ones the high-technology industries. His economic explanation of output preferences and related policies (1967*a*, pp. 6–9) drew upon the interest-group mechanism. Producer groups are the winners when the nationalistic ideology fosters a preference for placing status-giving jobs or property in the hands of nationals (see Breton 1964). This goal can be sought either by shifting the composition of output to create the preferred jobs or property (e.g., the preference for industrial output) or by removing these things from the hands of foreigners (restricting employment of foreign nationals or restricting the ownership of subsidiaries by foreign companies in order to place equity shares in domestic hands). These transfers serve both the collective consumption of nationalism and the domestic producer interests that benefit directly. They could also raise national economic welfare as conventionally measured if nationals were formerly discriminated against; but if less competent nationals get the same salaries as more competent foreigners, income is redistributed within the country.[5]

Johnson's thoughts on LDC nationalism probably grew from earlier bouts with the phenomenon in Canada. He had traced (1967*a*, chap. 6) a vicious circle in Canadian macroeconomic policies that began in the late 1950s, in which the key errors of commission and omission resulted from attempts to pursue nationalistic objectives with macroeconomic policy instruments. And he had attacked (1963, chap. 1) a Canadian proposal to curb the independence of foreign subsidiaries in Canada by promoting the sale of minority shareholdings to Canadian capitalists. Because the foreigners would still retain control of the companies, Johnson saw this proposal purely as service to local financial interest groups.

Other applications of behavioral theories of the political process yielded interesting if less sharp-edged results. The overall level of public-sector spending, he suggested (1975*b*, p. 25), might be inflated if majority coalitions succeed in loading part of the cost on unwilling minorities. But it might be suboptimal if the voters shun taxes to finance public goods. He concluded that some public-sector activities are probably overprovided (those "justified by a vague but persuasive national purpose"), others undersupplied. Johnson (1975*b*, chap. 13) found the political acquiescence in inflation somewhat resistant to an explanation from rational political behavior. A commitment by the government to maintain full employment locks it into asymmetrical situations that on balance generate inflation. But then why does the public vote a commitment to full employment that takes this form? Why does it give an election-bound government reason to think that a short-term bout of inflationary policy is likely on balance to win votes? An interest-group explanation did not satisfy Johnson; although the short-run losers from inflation (e.g., pensioners) may be politically weak, the general voter's contemplation of his permanent income and ultimate pensioner status cannot lead him indefinitely to support inflationary policies. Johnson held that informed public opinion under-stands the consequences of macroeconomic policy and indeed anticipates them, and he evidently could not countenance the suggestion that the public votes on the basis of a short time horizon or a biased valuation of the long-run potential relation between inflation and unemployment.[6]

---

[5] Nationalism also supports a preference for regulation of the economic activity carried on by alien entrepreneurs, even if it does not indicate exactly how the market outcome should be changed. Johnson wrote on several occasions about the clash between the nation-state, with this built-in procedural preference for control, and the multinational company able to pursue options outside of any given country (see Johnson 1967*a*, pp. 13–14; 1975*c*, chap. 5).

Occasionally Johnson's efforts to explain policies that fail to maximize real income led him to hint that voters' expressed preferences might give significant weight to objectives not counted in the economist's conventional view of utility maximization. Policies for assisting distressed regions tend to sustain factors of production in their traditional locations turning out marginal products smaller than they could in other regions. This might be the preference, stressed by Scitovsky (1976), for continuity rather than maximum income.[7] Also, the upsurge of concern for poverty in the 1960s, as in the 1930s, followed a period of sustained unemployment and could well reflect the feeling of the middle class that the poor might be doing badly relative to their own reasonable expectations, even if no worse off relative to the average income recipient (Johnson 1975*b*, pp. 5, 29–30, 220).

But if voters may violate the axioms of welfare economics, they can also act in the long run to avert the force of ill-conceived macroeconomic policies of central governments, or simply react to the differing regional incidence of these policies. Either consequence of national policy leads voters to appeal to lower and more politically sensitive levels of government, promoting the decentralization of federal systems and regional nationalism in countries with centralized political organs. This trend of voters' preferences is supported by an affluent society's increasing concern with extern [illegible] their incidence and, hence, more the pr [illegible] (Johnson 1975*b*, pp. 177–78).

Scattered through Johnson's writings [illegible] tal behavior patterns in international ec [illegible] are too much of their time to warrant rev [illegible] an emphasis that runs through them—th [illegible] national prestige with national interest (e.g. [illegible] International monetary relations then become [illegible] the contention of national rivalries." Because the pursuit of national power and prestige is a zero-sum game, its confusion with the pursuit of agreement on international economic relations (with their potential for positive-sum outcomes) is fundamentally divisive. Nations may negotiate institutional arrangements such as the Bretton Woods agreement that impose a politically agreeable but fictitious equality on members of the system, only to find (with the subsequent emergence of the dollar standard) that real differences in countries' economic situations determine what arrangements are actually viable. When national economic rationality does prevail in international economic relations, however, it tends to reach outcomes similar to those reached in individual bargaining. For example (1975*c*, p. 9), the free-rider problem explains governments' efforts to obtain for their citizens use of the modern technology without paying for the research that created it.

---

[6] He did flirt with the suggestion that the household division of labor involves an imperfect contract that motivates the husband to vote for full employment, the wife for lower prices.

[7] Even here, voters' choices may be rational, according to Johnson (1975*a*, p. [illegible]). The voter knows that political decisions are based on short time horizons and are always reversible; hence the preferences registered at the ballot box rationally reflect maximization over a short time horizon

III. The University and the Production of Knowledge

Johnson's eye turned frequently to social influences on the development of economic analysis and the quirks of the institutions that house its practitioners. A lifetime commuter among three countries, a compulsive guest at many houses of intellect, and a connoisseur of the swordplay of professional debate, he was well placed to spot the secrets of personal and institutional effectiveness in the processes of research and scholarship.

He set down one major statement about the university as a socioeconomic institution (1975*b*, chap. 11) and touched on several passing issues related to how many of society's resources it should command. The university he saw as a multiproduct firm, with some of its outputs public goods and most of them inappropriately priced. It has traditionally been under little pressure to maximize or even contemplate the productivity of its resource use. It provides public goods "by either setting standards of taste for the rest of the population or enabling the rest of the population to increase its productivity, command over consumption goods, or both" (1975*b*, p. 153). Qualitatively, this public-good problem is correctly being solved by public support and cost-free availability of the public goods (notably research), but the quantity is not necessarily optimized. Outputs of the university with increasingly less public character are the storage, organization, and synthesis of information, provision of general undergraduate education, and provision of professional education.

Johnson suggested a series of hypotheses about the economic function of general university education: (1) that it chiefly fosters the maturation of its students (it does not matter what you study, so long as it is rigorous and conveys cultural ambience); (2) that it filters out those best qualified for the "good jobs" that society has to offer (in which case the resources wasted educating those who pass through the filter pose a major question of efficiency); and (3) that its economic function is to create human capital with a discernible market value (so that it should attract not necessarily the best students but rather the most educable). The relative explanatory power of these hypotheses evidently has major implications for both the commitment of resources to universities and the character of training that they provide. Several factors discourage rational thinking about these problems of allocating resources in the university. The human-capital hypothesis faces a strong ideological opposition to the postulate that a university functions to whittle round pegs for society's round holes. And the attitude that universities should redistribute income to the young and intellectually ambitious beclouds the rational pricing of university education. University teachers, wrapped in the banner of the self-governing community of scholars, have dealt themselves terms of compensation that adduce an activity pattern poorly aligned with maximizing the value of the university's social product. Government finance of universities is blighted by information constraints that misallocate resources toward the more measurable inputs and outputs and occasionally distort the total amount on suspiciously transitory grounds (the faith in universities as progenitors of economic growth during the 1950s and 1960s). University administrations have little incentive to fight the faculty's misallocation of effort because the waste is not perceived in the outside world and because in most times opportunities for the administrators to pursue the bureaucratic quiet life are abundantly present.

Johnson's thoughts about a university's mostly set down against the backgroun 1960s. He was annoyed at university facul nize that the current generation of stude versity's "filter" function) had the incenti secure soft options and at their acquiesc university is a sovereign state with "oppresse classes (1972*a*). But he recognized (1968) among the causes of protest the mutually disappointed expectations that had flowed from the great expansion of university education. Heavily expanded and democratized universities had diluted (through expansion) the quality of their teaching staffs, and by greatly increasing the supply of graduates, they could no longer guarantee them a clear track into prestige jobs. Neither outcome was necessarily bad if properly adjusted to, but both created a disequilibrium of expectations.

Science policy, an issue of intellectual resource use related to the organization of universities, also attracted Johnson (1972*c*; 1975*c*, chap. 2). He found a good deal of muddy thinking to criticize. Procedures for evaluating the funding of applied research were being inappropriately applied to basic research and thereby creating an inducement for proponents of science to make excessive claims for the immediate payout of basic research. Nations were falling into confusion between scientific inputs and outputs, not recognizing that the international communication of scientific results frays the input-output relation within the individual nation's borders. The United States' scientific leadership has rested on "the capacity to put science and technology to work in the service of raising the standard of living of the masses, and the capacity to mobilise science and technology on a massive scale in the development of superior armaments" (1975*c*, p. 16), but this success is not clearly connected to public support of basic research. Scientists may argue for the support of an evolving "science culture," but they thereby make science a consumption good that must establish its claim against other social consumption goods. In sum, although the private market alone would underinvest in research, it is not clear that the actual market (a mix of government, private, and charitable actors) does so. And it is clear that the competitive progress of science, with a scientist's reward due to acclaim by fellow scientists, bears no necessary connection to a test of social usefulness.

Oddly enough, Johnson—whose personal needs for research inputs did not extend beyond pads of yellow paper—never applied these thoughts to the allocation of research resources to and within the economics profession.[8] The nearest he came was to address the question how far to extend programs of graduate education in economics (1973). These had proliferated during the 1960s (at least in Canada) beyond the optimal level. The proliferation had been promoted on the proposition that teaching and research interact. But this does not mean, he held, that they need to be carried on either by the same individual or at the same institution. He clearly felt that the research elite of the profession should be allowed to collect at leading institutions of graduate instruction, undistracted by undergraduates and free to enjoy the fruits of intellectual agglomeration and scale economies in graduate instruction. This might seem unfair to those sent down to teach the principles course, but it was an efficient use of resources.

---

[8] He wrote (1975*a*, p. 6). "I myself am fairly phlegmatic about and not much concerned with the question of whether economics is immediately socially useful in solving the current problems of society as society itself conceives them. . . . The current demand for 'relevance' . . . invariably amounts to an effort to force economists to lend their voices . . . to propaganda in support of the demander's political program "

Johnson wrote about both the influence of environmental conditions on the progress of economics and the forces governing the outcomes of debates within the profession. His travels among Canada, Britain, and the United States caused him to compare the effectiveness of their ambient conditions for nurturing scholarship (1975*b*, chap. 9). After World War II the United States had both the material resources and the aura of confidence to provide economists with a supportive environment. Also, scholars abroad reaching professorial status automatically become elite figures and score their points outside of intellectual endeavor, whereas the academic world of the United States is more self-contained and competitive. A pragmatic interest in research, boosted by the Sputnik boom in science, helped to consolidate and advance the American profession's lead. University tenure systems in the United States emphasize the requirement to "publish or perish," which disimpacts knowledge and discourages scholars from indulging in personal idiosyncrasy. The sheer size of the American economics profession allows a number of institutions to assemble critical masses of qualified people with serious research interests. Also (1975*b*, p. 151), it discourages scholars from "setting themselves up as across-the-board personal adversaries expounding conflicting holistic views of the subject" in favor of the modern economist's competition with the faceless host of his professional colleagues.

Anglo-American differences in political systems are also important, as is the general influence on economists' behavior
for the government. Ventures into polic
discourse into political rhetoric and cast
"as both interpreters of the political will
own interpretation of it" (1974, p. 8; see
Johnson recognized (1975*b*, pp. 177–79)
ment demand for economists had allowe
attain economies of scale and enlarged the
the future academic stars are drawn. The casual involvement
economists in government he found to be rather more benign and scientifically oriented than in Britain, where their public-sector work is less likely to lead to scholarly publication and more likely to cater to the current and prospective interests of one political party.

Johnson relied totally on the competition for esteem among highly skilled economists as the best guide for scientific progress, against any test that might be imposed from the outside. However, his faith in this nonmarket allocation mechanism did not blind him to the effects of the participants' finite human abilities and the moral hazard associated with scoring debating points. In various writings (1975*b*, chaps.

5–8) he reflected on the success of the Keynesian revolution and its challenge by the monetarist counterrevolution. In a scientific discipline an "orthodoxy" is necessary and desirable to facilitate communication in the profession and promote cumulative additions to knowledge, but it builds vested interests that resist any major overhaul (1975*b*, chap. 7). Both the Keynesian revolution and monetarist counterrevolution thus found their opportunities only when the prevailing orthodoxies had become unable to explain some facts of pressing importance. The would-be revolution then has to "attack the central proposition of conservative orthodoxy." It has to appear new and yet absorb enough of the valid elements of the prevailing orthodoxy to usurp the whole territory. It has to be hard enough to understand that young "comers" in the profession will invest in it in order to unseat more senior figures unable or unwilling to make the investment. It has to allow room for new or stylish methodologies of theoretical and empirical research. With evident glee Johnson analyzed how both Keynes and the monetarists had managed in their times to meet all these tests.

The challenge of radical economics also caught Johnson's eye (1977). Although intellectually unsympathetic, he took a detached view of what he saw as a failed attempt at an intellectual revolution. The interest of economists in social philosophy was driven undergound by McCarthyism, "the period of witch-hunting on behalf of cultural purity and loyalty that has twice followed involvement of an immigration-based new nation in a totalitarian-democratic foreign war" (1977, p. 99). The expansion of government employment for economists and the post-Sputnik faith of the public in science as panacea all maintained this suppression. And economics teaching, lacking a market test and drawn toward preprofessional subject matter, increasingly lost contact with the humanistic tradition. When the Vietnam War shifted the balance, the scientific cast of standard research and training in economics both weakened and strengthened the traditionalists' response to the radical challenge. It was strengthened because the Keynesian and post-Keynesian experience in debating emotionally charged issues had instilled the attitude that conflicting ideas should be turned into testable propositions. But it was weakened by a habitual assumption that the protesters "were scientifically rational and maturely sincere in their protest." In any case, the radicals did not meet the test for a successful intellectual revolution, and Johnson saw its members as being ingested into the profession as they sought out viable research options in economic history and the history of thought.

The great bulk of Johnson's writing in social science is positive or only tentatively normative, but he did not conceal his views about the appropriate conduct of economic research itself. He applauded the increasing acceptance of the competitive model and pragmatism in the treatment of values as yielding increased professional competence and confidence (1975*b*, pp. 19–20). Given a diverse scientific community that judges the value of research projects and the quality of research design and execution, there is no great need for concern that partisanship will influence the selection of projects or result from the findings of scientifically conceived research (1975*b*, pp. 140–42). "Much of the work of economists concerned with policy issues is devoted to sorting out the true ends of policy from the means intended to achieve these ends, and to assessing the relevance and relative efficiencies of the various means proposed—in short, to determining what the problem really is, and attempting to evaluate the various ways of solving it" (1975*b*, p. 18).

References

Breton, Albert. "The Economics of Nationalism" *J P E* 72, no 4 (August 1964): 376–86.

Downs, Anthony. *An Economic Theory of Democracy*. New York: Harper, 1957

Galbraith, John Kenneth. *The Affluent Society* Boston. Houghton Mifflin. 1958

Johnson, Harry G. *The Canadian Quandary Economic Problems and Policies* Toronto: McGraw-Hill, 1963

———. "An Economic Theory of Protectionism, Tariff Bargaining, and the Formation of Customs Unions" *J P E* 73, no 3 (June 1965) 256–83

———, ed. *Economic Nationalism in Old and New States* Chicago Univ Chicago Press, 1967. (*a*)

———. "Some Economic Aspects of Brain D
(Autumn 1967). 379–411. (*b*)

———. "The Economics of Student Pro
1968): 673–75

———. "The Dilution of Academic Pow
1972): 486–90. (*a*)

———. "The International Monetary Syst
and Econ. 15 (October 1972). 277–92. (*b*

———. "Some Economic Aspects of Science *M*
18. (*c*)

———. "The Uneasy Case for Universal Graduate Programmes in Economics." *Minerva* 11 (April 1973). 263–68

———. "Individual and Collective Choice" In *Man and the Social Sciences*, edited by William A. Robson. Beverly Hills, Calif. Sage, 1974

———. "Equity and Economic Theory." *Nebraska J Econ and Bus* 14 (Summer 1975): 3–17 (*a*)

———. *On Economics and Society* Chicago: Univ Chicago Press, 1975 (*b*)

———. *Technology and Economic Interdependence* London. Macmillan (for Trade Policy Res. Centre), 1975 (*c*)

———. "Economics and the Radical Challenge: The Hard Social Science and the Soft Social Reality." In *Culture and Its Creators: Essays in Honor of Edward Shils*, edited by Joseph Ben-David and Terry N. Clark. Chicago: Univ. Chicago Press, 1977.

Krueger, Anne O. "The Political Economy of the Rent-seeking Society." *A.E.R.* 64 (June 1974): 291–303.

Reuber, Grant L., and Scott, Anthony D. "Harry Gordon Johnson, 1923–1977." *Canadian J. Econ.* 10 (November 1977): 670–77.

Scitovsky, Tibor. *The Joyless Economy: An Inquiry into Human Satisfaction and Consumer Dissatisfaction.* New York: Oxford Univ. Press, 1976.

Harry G. Johnson: A Bibliography

Vicky M. Longawa, Editor and Compiler
University of Chicago

A. Books and Lecture Series

1. 1952. *The Overloaded Economy: The Economic Problems of Great Britain.* Toronto: Univ. Toronto Press.
2. 1958. *International Trade and Economic Growth. Studies in Pure Theory.* London: Allen & Unwin; Cambridge, Mass.: Harvard Univ. Press. 2d printing, 1967. Translations: 1960 (Japanese, with new preface); 1970 (Spanish); 1973 (Portuguese); 1976 (Italian).
3. 1962. *Canada in a Changing World Economy.* Alan B. Plaunt Memorial Lectures, Carleton University. Toronto: Univ. Toronto Press.
4. 1962. *Lags in the Effects of Monetary Policy in Canada* (with John W. L. Winder). Ottawa: Queen's Printer (for Royal Commission on Banking and Finance).
5. 1962. *Money, Trade and Economic*
nomic Theory. London: Allen & U
vard Univ. Press. 2d ed., 1964. R
1964 (Japanese, with new prefac
6. 1963. *The Canadian Quandary: F*

The bibliography was originally compiled by Johnson's various secretaries, especially Patricia Hume. My work involved checking the items for accuracy by going back to original sources where possible, inserting any items that were left out (Johnson did not list many reprints of his articles, quite a few of which I found by accident), and arranging the entire list chronologically. This bibliography cannot be considered "complete": the diversity of subject matter (as well as the number of obscure foreign sources) renders nearly impossible the task of checking all possibilities. E. J. Belton, of Lakehead University in Thunder Bay, Ontario, compiled another bibliography based more heavily on Johnson's original list.

[*Journal of Political Economy*, 1984, vol. 92, no. 4]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9204-0010$01.50

ronto: McGraw-Hill. Reprint ed., Toronto: McClelland & Stewart, Carleton Library, 1977.

7. 1965. *Trade and Development* (with Peter B. Kenen). Geneva: Librairie Droz.
8. 1965. *The World Economy at the Crossroads.* K. G. Norris Memorial Lectures. Montreal: Canadian Trade Commission; New York: Oxford Univ. Press; Oxford: Clarendon. 2d impression (with new preface), 1968. Translations: 1967 (Japanese); 1967 (Swedish); 1968 (Spanish, Barcelona: Editorial Labor); 1969 (Danish).
9. 1967. *Economic Policies toward Less Developed Countries.* Washington: Brookings Inst.; London: Allen & Unwin; New York: Praeger, 1968. Translations: 1970 (Spanish, Santiago: Univ. Catolica de Chile); 1972 (Japanese).
10. 1967. *Essays in Monetary Economics.* London: Allen & Unwin; Cambridge, Mass.: Harvard Univ. Press. 2d ed., 1969. Translations: 1969 (German, Berlin: Duncker & Humblot); 1971 (Spanish).
11. 1967. *Trade and Aid Policies: The UNCTAD Alternatives.* Washington: Brookings Inst.
12. 1968. *Comparative Cost and Commercial Policy Theory for a Developing World Economy.* The Wicksell Lectures. Stockholm: Almqvist & Wiksell. Reprinted in *Pakistan Development Rev.* 9 (suppl.; Spring 1969): 1–33. Also in [A.24].[1]
13. 1968. *Harmonization of National Economic Policies under Free Trade* (with Paul Wonnacott and Hirofumi Shibata). Toronto: Univ. Toronto Press.
14. 1969. *Rebuilding the Liberal Order* (with D. H. Barron and the Earl of Cromer). London: Inst. Econ. Affairs.
15. 1971. *Aspects of the Theory of Tariffs.* London: Allen & Unwin; Cambridge, Mass.: Harvard Univ. Press, 1972.
16. 1971. *Macroeconomics and Monetary Theory* London: Gray-Mills; Chicago: Aldine, 1972.
17. 1971. *The Two-Sector Model of General Equilibrium.* Yrjo Jahnsson Lectures. London: Allen & Unwin; Chicago: Aldine.
18. 1972. *Further Essays in Monetary Economics.* London: Allen & Unwin; Cambridge, Mass.: Harvard Univ. Press, 1973. Translation: 1976 (German, Berlin: Duncker & Humblot).
19. 1972. *Inflation and the Monetarist Controversy.* F. De Vries Lectures. Amsterdam and London: North-Holland. Reprint ed., 1976. Also in [A.23].
20. 1973. *The Theory of Income Distribution.* London: Gray-Mills.

---

[1] Items in brackets refer to Johnson's books listed in Sec. A or C in which an article has been reprinted.

21. 1974. *General Equilibrium Analysis: A Microeconomic Text* (with Melvyn B. Krauss). London: Allen & Unwin; Chicago: Aldine, 1975.
22. 1975. *Money, Balance-of-Payments Theory, and the International Monetary Problem.* David Horowitz Lectures, Israel. In *Bangladesh Development Studies* 4 (October 1976): 455–78. Essays in International Finance, no. 124. Princeton, N.J.: Princeton Univ. Press, 1977. Also in *International Economic Policies and Their Theoretical Foundations: A Source Book,* edited by John M. Letiche. New York: Academic Press, 1982.
23. 1975. *On Economics and Society.* Chicago and London: Univ. Chicago Press; Taipei, Taiwan: Ch'eng Wen, 1976.
24. 1975. *Technology and Economic Interdependence.* London: Macmillan (for Trade Policy Research Centre); Toronto: Macmillan of Canada, 1976; New York: St. Martin's. Translation: 1978 (Spanish, Mexico City: Editoral el Manual Moderno).
25. 1976. *Trade Negotiations and the New International Monetary System.* Geneva: Grad. Inst. Internat. Studies.
26. 1978. *Selected Essays in Monetary Economics.* London and Boston: Allen & Unwin.
27. 1978. *The Shadow of Keynes: Understanding Keynes, Cambridge, and Keynesian Economics* (with Elizabeth S. Johnson). Chicago: Univ. Chicago Press; Oxford: Blackwell.

B. Pamphlets

1. 1941. *The Antigonish Movement: A*
   University. Antigonish, Nova Scotia:
2. 1961. *The Social Policy of an Opulent*
   Ottawa: Canadian Welfare Council
3. 1963. *Alternative Guiding Principles*
   Essays in International Finance, no
   ton Univ. Press, 1963. Also in *Canadian*
   Policy, edited by James P. Cairns and H. H Binhammer. Toronto: McGraw-Hill, 1965. Also in [A.6; A.10]
4. 1963. *Economic Growth and Economic Policy.* J. M. McLean Lecture Guelph: Ontario Agriculture College. Also in [A.6]
5. 1967. *The Economic Approach to Social Questions.* Inaugural lecture, London School of Economics, October. London: Weidenfeld & Nicolson, 1968. Also in *Economica* 35 (February 1968): 1–21. Also in *Public Interest,* no. 12 (Summer 1968), pp. 68–79. Also in *Mercurio* 11 (September 1968): 1–6. Also as "Economic Concepts and Social Questions." In *Understanding Economics: Essays on Public Policy,* edited by Yung-Peng Chen. Boston: Little, Brown, 1974. Also in [A.23].

6. 1970. *The "Crisis of Aid" and the Pearson Report.* Lecture presented at Edinburgh University. Edinburgh: Edinburgh Univ. Also in *Foreign Resources and Economic Development,* edited by T. J. Byres. London: Cass, 1972. Also in *International Economics and Development,* edited by Luis E. DiMarcos. New York and London: Academic Press, 1972. Also in Spanish in *Economia internacional y desarolla.* Buenos Aires: Depalma, 1974.
7. 1970. *Efficiency in Domestic and International Money Supply.* International Economics no. 3. Guildford: Univ. Surrey.
8. 1970. *An Overall View of International Economic Questions Facing Britain, the United States, and Canada during the 1970s.* London: British–North American Comm.
9. 1972. *Memorial to the Prime Minister* (with A. A. Walters et al.). London: Econ. Radicals.
10. 1973. *Man and His Environment.* London: British–North American Comm. Also in [A.23; A.24].
11. 1973. *The Problem of International Monetary Reform.* Stamp Memorial Lecture. London: Athlone, 1974.
12. 1974. *Canadian Relations with the United States.* Proceedings of the Standing Senate Committee on Foreign Affairs, issue no. 3. Ottawa: Senate of Canada.
13. 1974. *Dear Prime Minister* (with A. A. Walters et al.). London: Econ. Radicals.
14. 1975. *Economics for the Business Man.* London: Found. Bus. Responsibilities.

C. Edited Books

1. 1962. *International Monetary Arrangements: The Problem of Choice.* Princeton, N.J.: Princeton Univ. Press.
2. 1962. *Landmarks in Political Economy* (with Earl J. Hamilton and Al Rees). Chicago: Univ. Chicago Press.
3. 1965. *The Balance of Payments Statistics of the United States: A Review and Appraisal.* Washington: Government Printing Office.
4. 1965. *Trade, Growth, and the Balance of Payments: Essays in Honor of Gottfried Haberler* (with Richard E. Caves and Peter B. Kenen). Chicago: Rand McNally; Amsterdam, North-Holland.
5. 1966. *OECD Reviews of National Science Policy: France.* Paris: OECD.
6. 1967. *Economic Nationalism in Old and New States.* Chicago: Univ. Chicago Press; London: Allen & Unwin, 1968.
7. 1967. *Issues in Monetary Research 1966.* 1966 Conference of the American Bankers Assoc. Supplement to *J.P.E.*, vol. 75, no. 4, pt. 2 (August).

8. 1968. *Issues in Monetary Research 1967.* 1967 Conference of University Professors sponsored by the American Bankers Assoc. Supplement to *J.P.E.*, vol. 76, no. 4, pt. 2 (July/August).
9. 1968. *Readings in International Economics* (with Richard E. Caves). Homewood, Ill.: Irwin (for American Econ. Assoc.); London: Allen & Unwin.
10. 1969. *Issues in Monetary Research 1968.* 1968 Conference of University Professors sponsored by the American Bankers Assoc. Supplement to *J. Money, Credit and Banking*, vol. 1 (August).
11. 1969. *New Trade Strategy for the World Economy.* London: Allen & Unwin; Toronto: Univ. Toronto Press.
12. 1970. *Money in Britain, 1959–69: The Papers of the Radcliffe Report Ten Years After* (with David R. Croome). London: Oxford Univ. Press. Translation: 1973 (Italian).
13. 1971. *The Current Inflation* (with A. Robert Nobay). London: Macmillan; New York: St. Martin's.
14. 1971. *Effective Tariff Protection* (with Herbert G. Grubel). Geneva: Grad. Inst. Internat. Studies.
15. 1971. *Trade and Development: Essays in Economics by V. K. Ramaswami* (with Jagdish N. Bhagwati and T. N. Srinivasan). London: Allen & Unwin.
16. 1971. *Trade Strategy for Rich and Poor Nations.* London: Allen & Unwin; Toronto: Univ. Toronto Press.
17. 1972. *Non-Tariff Factors in Tariff-free Trade.* London: Allen & Unwin.
18. 1972. *Readings in British Monetary F*
19. 1972. *Towards an Open World E*
Macmillan (for Trade Policy R
tin's.
20. 1973. *The Daily Economist* (with B
Cliffs, N.J.: Prentice-Hall.
21. 1973. *The Economics of Common* (
Swoboda). London: Allen & Unwin; C
Univ. Press.
22. 1974. *Issues in Monetary Economics* (with A. Robert Nobay). London: Oxford Univ. Press.
23. 1974. *The New Mercantilism: Some Problems in International Trade, Money, and Investment.* Oxford: Blackwell; New York: St. Martin's.
24. 1976. *The Monetary Approach to the Balance of Payments* (with Jacob A. Frenkel). London: Allen & Unwin; Toronto: Univ. Toronto Press. Paperback: Allen & Unwin, 1977.
25. 1978. *The Economics of Exchange Rates: Selected Studies* (with Jacob A. Frenkel). Reading, Mass.: Addison-Wesley.

D. Scientific Articles

1948. "An Error in Ricardo's Exposition of His Theory of Rent." *Q.J.E.* 62 (November): 792–93.

1949. "Demand for Commodities Is Not Demand for Labour." *Econ. J.* 59 (December): 531–36.

———. "Malthus on the High Price of Provisions." *Canadian J. Econ. and Polit. Sci.* 15 (May): 190–202. Also in *Outside Readings in Economics*, edited by Arleigh P. Hess et al. New York: Crowell, 1951.

1950. "The Case for Increasing the Price of Gold in Terms of All Currencies: A Contrary View." *Canadian J. Econ. and Polit. Sci.* 16 (May): 199–209.

———. "The De-stabilising Effect of International Commodity Agreements on the Prices of Primary Products." *Econ. J.* 60 (September): 626–29.

———. "Diagrammatic Analysis of Income Variations and the Balance of Payments." *Q.J E.* 64 (November): 623–32.

1950. "Unrequited Imports and the Terms of Trade: A Note" (with Charles F. Carter). *Econ. J.* 60 (December): 837–39.

1950–51. "A Three-dimensional Model of the Shackle ϕ-Surface." *Rev. Econ. Studies* 18 (2): 115–18.

1951. "Clearing Bank Holdings of Public Debt, 1930–1950." *London and Cambridge Econ. Service* 29 (November): 102–9.

———. "A Note on the Effect of Income Redistribution on Aggregate Consumption with Interdependent Consumer Preferences." *Economica*, N.S. 18 (August): 295–97.

———. "Recent Trends in British Banking." *Company Accountant*, N.S. 14 (June): 10–14.

———. "Some Implications of Secular Changes in Bank Assets and Liabilities in Great Britain." *Econ. J.* 61 (September): 544–61.

———. "The Taxonomic Approach to Economic Policy." *Econ. J.* 61 (December): 812–32.

1951–1952. "Optimum Welfare and Maximum Revenue Tariffs." *Rev. Econ. Studies* 19 (1): 28–35. As "Alternative Optimum Tariff Formulae" in [A.2].

———. "Some Cambridge Controversies in Monetary Theory." *Rev. Econ. Studies* 19 (2): 90–104.

1952. "The Effects of Income Redistribution on Aggregate Consumption with Interdependence of Consumers' Preferences." *Economica*, N.S. 19 (May): 131–47.

———. "Lord Keynes and Modern Economic Thought." *Canadian Forum* 32 (November): 176–77.

———. "The Matrix Multiplier and an Ambiguity in the Keynesian Concept of Saving." *Econ. J.* 62 (March): 197–200.

———. "[Monetary Policy Again] Concluding Comment." *Bull. Oxford Univ. Inst. Statis.* 14 (August): 298–306.

———. "The New Monetary Policy and the Problem of Credit Control." *Bull. Oxford Univ. Inst. Statis.* 14 (April/May): 117–31.

1953. "Equilibrium Growth in an International Economy." *Canadian J. Econ. and Polit. Sci.* 19 (November): 478–500. Also in [A.2]. Also in German in *Wachstum und Entwicklung der Wirtschaft*, edited by Heinz Konig. Cologne: Kiepenheuer & Witsch, 1968.

———. "Interdependence of Consumer Preference in the Theory of Income Redistribution: Comment." *Econ. J.* 63 (March): 83.

———. "Recent Developments in British Monetary Policy." *A.E.R. Papers and Proc.* 43 (May): 19–26.

———. "The Reparations Problem: A Correction." *Econ. J.* 63 (September): 724–25.

1954. "Canada—a Lost Opportunity?" *Three Banks Rev.* 22 (June): 3–21.

———. "Increasing Productivity, Income-Price Trends and the Trade Balance." *Econ. J.* 64 (September): 462–85. Also in [A.2].

———. "The Macroeconomics of Income Redistribution." In *Income Redistribution and Social Policy*, edited by Alan T. Peacock. London: Jonathan Cape. Also in German in *Soziale Sicherheit*, edited by Bernhard Külp and Wilfrid Schreiber. Cologne: Kiepenheuer & Witsch, 1971.

———. "Optimum Tariffs and Retaliation." *Rev. Econ. Studies* 21 (2): 142–53. Also in [A.2]. Also in *Readings in Welfare Economics: A Selection of Papers from the Review of*
Farrell. London: Macmillan, 1973.

1955. "Are We Exporting Too Much

———. "Economic Expansion and In
School Econ. and Soc. Studies 23 (May
Expansion of World Trade and the
Significant Papers, edited by Richard
per & Row, 1968.

———. "Economic Expansion and the Balance of Payments." *Bull. Oxford Univ. Inst. Statis.* 17 (February): 3–10.

———. "Keynes and Supply Functions: A Mathematical Appendix" (to article by Dennis Robertson). *Econ. J.* 65 (September): 177–78.

———. "The Transfer Problem: A Note on Criteria for Changes in the Terms of Trade." *Economica*, N.S. 22 (May): 113–21.

1956. "A Further Note on Dynamic Economics." *Rev. Econ. Studies* 23 (3): 245–46.

———. "General Equilibrium Analysis of Excise Taxes: Comment." *A.E.R.* 46 (March): 151–56.

———. "Lecture 1: Banking and Monetary Policy"; "Lecture 2: The

Balance of Payments"; "Lecture 3: Comparative Costs and Commercial Policy"; "Lecture 4: Planning and the Market in Economic Development"; "Lecture 5: Monetary Theory and Keynesian Economics." Five lectures delivered to Pakistan Econ. Assoc. Refresher Course. *Pakistan Econ. J.* 8 (June 1958): 1–70. Also in [A.5], omitting Lecture 3. Lecture 5 also in *Monetary Theory: Selected Readings,* edited by Robert W. Clower. Baltimore: Penguin, 1969. Lecture 4 also as "The Market Mechanism as an Instrument of Development." In *Leading Issues in Economic Development,* edited by Gerald M. Meier. New York: Oxford Univ. Press, 1964; 2d ed., 1970; 3d ed., 1976; 4th ed., 1984. Also in *Pricing Policy for Development Management,* edited by Gerald M. Meier. Baltimore: Johns Hopkins Univ. Press (for Econ. Development Inst. of the World Bank), 1983.

———. "On British Crises." *Granta* 59 (April): 5–6.

———. "The Revival of Monetary Policy in Britain." *Three Banks Rev.* 30 (June): 3–20.

———. "A Simplification of Multi-Country Multiplier Theory." *Canadian J. Econ. and Polit. Sci.* 22 (May): 244–46. Also in [A.2].

———. "Sketch of a Generalization of Keynesian Balance-of-Payments Theory." *Indian J. Econ.* 37 (July): 49–56.

———. "The Transfer Problem and Exchange Stability." *J.P.E.* 64 (June): 212–25. Also in [A.2; C.9]. Also in *International Finance: Selected Readings,* edited by Richard N. Cooper. Harmondsworth: Penguin, 1969.

———. "Un versione diagrammatica della teoria dell'interesse dei fondi prestabili" [A diagrammatic version of the loanable funds theory]. *Riv. internaz. sci. econ. e commerciali* 3 (January): 16–23.

1957. "Bank Rate Reform and the Improvement of Monetary Statistics." *Bull. Oxford Univ. Inst. Statis.* 19 (November): 341–45.

———. "The Criteria of Economic Advantage." *Bull. Oxford Univ. Inst. Statis.* 19 (February): 33–38. Also in *The Free Trade Proposals,* edited by G. D. N. Worswick. Oxford: Blackwell, 1960.

———. "The Determination of the General Level of Wage Rates." In *The Theory of Wage Determination,* edited by John T. Dunlop. London: Macmillan.

———. "Discriminatory Tariff Reduction: A Marshallian Analysis." *Indian J. Econ.* 38 (July): 39–47.

———. "The Economics of Governor Stevenson's Program Paper: Where Is the Money Coming From?" *Rev. Econ. and Statis.* 39 (May): 139–41.

———. "The European Common Market: Risk or Opportunity? A British View." *Weltwirtschaftliches Archiv* 79 (2): 267–82.

———. "Factor Endowments, International Trade and Factor Prices."

Manchester School Econ. and Soc. Studies 25 (September): 270–83. Also in [A.2; C.9].

———. "Notes on Economic Development and the Maximum Rate of Growth." *Malayan Econ. Rev.* 2 (April): 16–22.

1958. "Canada's Economic Prospects." *Canadian J. Econ. and Polit. Sci.* 24 (February): 104–10. Also in [A.6].

———. "A Consistent Set of Principles for the Conduct of Domestic and Foreign Economic Policy." In *Problems of the United States Economic Development*, by the Committee on Economic Development. New York: Comm. Econ. Development.

———. "Demand Theory Further Revised, or Goods Are Goods." *Economica*, N.S. 25 (May): 149.

———. "The Economic Gains from Freer Trade with Europe." *Three Banks Rev.* 39 (September): 3–19.

———. "The Economic Prospects." *Listener* 59 (January 9): 47–48

———. "The Gains from Freer Trade with Europe: An Estimate." *Manchester School Econ. and Soc. Studies* 26 (September): 247–55. Also in *The Free Trade Proposals*, edited by G. D. N. Worswick. Oxford: Blackwell, 1960.

———. "In the Event of a Slump." *District Bank Rev.*, no. 126 (June), pp. 18–29.

———. "Marshallian Analysis of Discriminatory Tariff Reduction. An Extension." *Indian J. Econ.* 39 (October): 177–81.

———. "Notes on the Theory of Optimum Trade Restrictions." *Pakistan Econ. J.* 8 (September): 16–52.

———. "Sales Taxes and Import Taxes: A Comment" (with William Peters). *De Economist* 6 (June): 476–79.

———. "Tariffs and the Balance of Paym
(May): 301–2.

———. "Towards a General Theory
[A.2]. Also in French in *Cahiers d*
1959): 1–12. Also in [C.6; C.9]. Als
Readings, edited by Richard N. Coo
1969. Also as "A General Theory of
Macroeconomic Themes: Edited Readings in M
M. J. C. Surrey. London: Oxford Univ. Press, 1976. Also in [C.24].

———. "Two Schools of Thought on Wage Inflation." *Scottish J. Polit. Econ.* 5 (June): 149–53.

1958–59. "Economics of Freer European Trade." *Manchester Chamber of Commerce Directory.*

1959. "British Monetary Statistics." *Economica*, N.S. 26 (February): 1–17.

———. "Convertibility and After." *Canadian Forum* 38 (March): 271–72.

———. "Economic Development and International Trade." *Pakistan Econ. J.* 9 (December): 47–71. Also in *Nationalokonomisk Tidsskrift* 97 (5, 6): 253–72. Also in [A.5; C.6; C.9].

———. "The Index-Number Theorem: A Geometrical Note." *Canadian J. Econ. and Polit. Sci.* 25 (May): 174–79.

———. "International Trade, Income Distribution, and the Offer Curve." *Manchester School Econ. and Soc. Studies* 27 (September): 241–60. Also in [A.15].

1960. "Appraisal of the Far Eastern and Latin American Team Reports in the World Foreign Trade Setting." In *Increasing Understanding of Public Problems and Policies.* Chicago: Farm Found.

———. "Arthur Cecil Pigou, 1877–1959" (obituary). *Canadian J. Econ. and Polit. Sci.* 26 (February): 150–55. Also in [A.27].

———. "Canada's Foreign Trade Problems." *Internat. J.* 15 (Summer): 233–41. Also in [A.6].

———. "The Common Market: The Economists' Reaction." In *The Free Trade Proposals*, edited by G. D. N. Worswick. Oxford: Blackwell.

———. "The Consumer and Madison Avenue." *Current Econ. Comment* 22 (August): 3–10.

———. "The Cost of Protection and the Scientific Tariff." *J.P.E.* 68, no. 4 (August): 327–45. Also in [A.15].

———. "The Economic Theory of Customs Union." *Pakistan Econ. J.* 10 (March): 14–32. Also in [A.5].

———. "Extract of Evidence." In *Minutes of Evidence*, by the Committee on the Working of the Monetary System. London: H.M.S.O.

——— "Francis Ysidro Edgeworth." *Handwörterbuch der Sozialwissenschaften* 28 (January): 25–26.

———. "Income Distribution, the Offer Curve, and the Effects of Tariffs." *Manchester School Econ. and Soc. Studies* 28 (September): 215–42. Also in [A.15].

———. "Memorandum of Evidence." In *Principal Memoranda of Evidence*, vol. 3, by the Committee on the Working of the Monetary System. London: H.M.S.O.

———. "Notes on Some Controversies in the Theory of International Trade" (with Jagdish N. Bhagwati). *Econ. J.* 70 (March): 74–93.

———. "The Political Economy of Opulence." *Canadian J. Econ. and Polit. Sci.* 26 (November): 552–64. Also in [A.5; A.6; A.23].

———. "Les Problèmes monétaires dans un zone de libre échange." In *Marché commun, institutions communes*, edited by U. Arangio-Ruiz. Paris: Pichon & Durand-Auzias.

———. "The Pure Theory of International Trade: Comment." *A.E.R.* 50 (September): 721–22.

———. "Symposium on the Report of the Trade and Tariffs Commission (Croft Report): Comment." *Soc. and Econ. Studies* 9 (March): 1–12.

1961. "The Case for Tariff Reform." *Bus. Q.* 26 (Spring): 25–33. Also in [A.6].

———. "Economic Implications of the Size of Countries." *Econ. Development and Cultural Change* 10 (October): 105–8.

———. "Effects of Changes in Comparative Costs as Influenced by Technical Change." *Malayan Econ. Rev.* 6 (October): 1–13. Also in Italian in *Econ. internaz.* 15 (August/November 1962): 113–32. Also in *International Trade Theory in a Developing World*, edited by Roy Harrod and Douglas C. Hague. London: Macmillan, 1964. Also in French in *Cahiers de l'Inst. Sci. Econ. Appl.*, vol. 87 (November 1965).

———. "External Economics and International Trade: Comment." *Osaka Econ. Papers* 9 (March): 22–27.

———. "Factor Mobility and Rates of Technical Progress: A Critical Note on 'Output, Labour and Capital in the Canadian Economy.'" *Canadian J. Econ. and Polit. Sci.* 27 (February): 101–5.

———. "A Generalized Theory of the Effects of Tariffs on the Terms of Trade" (with Jagdish N. Bhagwati). *Oxford Econ. Papers*, N.S. 13 (October): 225–53. Also in [A.15].

———. "The *General Theory* after Twenty-five Years." *A.E.R. Papers and Proc.* 51 (May): 1–17. Also as "Keynes' *General Theory* after Twenty-five Years" (with notes; English/Chinese). *Hong Kong Econ Papers* 1 (June): 33–51. Also in [A.5; A 23].

———. "The International Liquidity
ments Imbalances and Need for St
Arrangements. Hearings before the
Exchange and Payments of Joint
gress, May 16 and June 19–21,
Printing Office.

———. "The International Perspective
Free Trade Areas." In *Canada and the*
Three Essays, edited by Harry E. English. Toronto: Univ. Toronto Press. Also in [A.6].

———. "Notes on Some Controversies in the Theory of International Trade: A Rejoinder" (with Jagdish N. Bhagwati) *Econ J* 71 (June): 427–30.

———. "Problems of Canadian Nationalism." *Internat. J* 16 (Summer): 238–49. Also in [A.6].

1962. "Canadian-American Integration in Face of a Changing World Economy." In *Fourth Seminar on Canadian-American Relations. Proceedings*, edited by F. J. Boland. Windsor, Ont.: Assumption Univ. Windsor Press. Also in [A.6].

———. "Canadian-American Trade and the Common Market." *Bus. Topics* 10 (Spring): 39–46.

———. "The Challenge to Canada: Excerpts from a Forum" (with M. Wardell and R. M. Fowler). In *Canada, the Commonwealth and the Common Market*, edited by William B. Cunningham. Montreal: McGill Univ. Press.

———. "International Liquidity: Problems and Plans." *Malayan Econ. Rev.* 7 (April): 1–19. Also in [A.6].

———. "International Trade and Economic Growth: A Supplementary Analysis." *Arthaniti* 5 (July): 140–52.

———. "Monetary Theory and Policy." *A.E.R.* 52 (June): 335–84. Also in *Surveys of Economic Theory*, vol. 1. London: Macmillan; New York: St. Martin's (for American Econ. Assoc.), 1965. Also in *Monetary Theory and Policy: Major Contributions to Contemporary Thought*, edited by Richard S. Thorn. New York: Random House, 1966. Rev. ed., New York: Praeger, 1976. Also as "The Objectives and Instrumental Role of Monetary Policy." In *Money and Finance: Readings in Theory, Policy, and Institutions*, edited by Deane Carson. New York: Wiley, 1966. Also in [A.10].

———. "A Simple Joan Robinson Model of Accumulation with One Technique." *Osaka Econ. Papers* 10 (February): 28–33.

———. "The Unemployment Problem." In [A.3]. Also in *Canadian Banking and Monetary Policy*, edited by James P. Cairns and H. H. Binhammer. Toronto: McGraw-Hill, 1965.

———. "Where the World Is Going." In *Canada, the Commonwealth and the Common Market*, edited by William B. Cunningham. Montreal: McGill Univ. Press, 1962. Also in [A.6].

1962–63. "Canada in a Changing World." *Internat. J.* 18 (Winter): 17–28. Also in [A.6].

1963. "Advertising in Today's Economy." *Queen's Q.* 69 (Winter): 537–54. Also in [A.6].

———. "An Appraisal of Kennedy's International Economic Policies." *Bus. Horizons* 6 (Summer): 81–86.

———. "The Bladen Plan for Increased Protection of the Canadian Automotive Industry." *Canadian J. Econ. and Polit. Sci.* 29 (May): 212–38. Also in [A.6; A.15].

———. "The Bladen Plan: A Reply." *Canadian J. Econ. and Polit. Sci.* 29 (November): 515–18.

———. "Equilibrium under Fixed Exchanges." *A.E.R. Papers and Proc.* 53 (May): 112–19.

———. "Essentials of Economic Development." *Econ. Development and Cultural Change* 12 (October): 94–98.

———. "Lags in the Effects of Monetary Policy in Canada." In [A.6]. Also in French in *L'Actualité écon.* 39 (July/September): 175–98.

———. "Notes on the Theory of Transactions Demand for Cash." *Indian J. Econ.* 44 (July): 1–11. Also in [A.10].

———. "An Objective Analysis of the Objectives of American Monetary Policy, 1952–61" (with William G. Dewald). In *Banking and Monetary Studies,* edited by Deane Carson. Homewood, Ill.: Irwin.

———. "Objectives, Monetary Standards, and Potentialities." In *The State of Monetary Economics,* conference of the Universities–National Bureau Committee on Economic Research. *Rev. Econ. and Stats.* 45 (suppl.; February): 137–44.

———. "An Overview of Price Levels, Employment, and the Balance of Payments." *J. Bus.* 36 (July): 279–89. Also in [A.6]. Also in *The Dollar Deficit: Causes and Cures,* edited by H. Peter Gray. Boston: Heath, 1967.

———. "Planning and the Objectives of a Free Enterprise Economy." Paper presented at the Ninth Winter Conference of CIPA, Toronto, February. In *Economic Planning in a Democratic Society?* edited by T. E. H. Reid. Toronto: Univ. Toronto Press (for Canadian Inst. Public Affairs). Also in [A.6].

———. "Recent Developments in Monetary Theory." *Indian Econ Rev.* 6 (August): 1–28. Also in Italian in *Quaderni di sci. soc.* 5 (April): 39–68. Also in [A.10].

———. "Stefan Stykolt, 1923–1962" (obituary). *Canadian J. Econ. and Polit. Sci.* 29 (February): 99–104.

———. "A Survey of Theories of Inflation." *Indian Econ. Rev.* 6 (August): 29–69. Also in Italian in *Quaderni di sci. soc.* 3 (December 1964): 354–90. Also in [A.10].

———. "Towards a Generalized Ca
Economic Development." In [A.
Economic Growth. Paris: OECD, 19
vol. 1, *Selected Readings,* edited by
1968. Also in *Economic Developmen*
edited by Theodore Morgan and
Wadsworth, 1970. Also as "Develop
of Capital Accumulation." In *Leading Issues*
4th ed., edited by Gerald M. Meier. New York: Oxford Univ. Press
1984.

1964. Comments on John Vaizey's "Towards a New Political Economy? or Some Problems of Some Aspects of Economics in the Light of 'Human Resource' Concepts." In *The Residual Factor and Economic Growth.* Paris: OECD.

———. "Economic Growth as an Objective of Government Policy: Discussion." *A.E.R. Papers and Proc.* 54 (May): 21–24.

———. "Economic Nationalism in Canadian Policy." *Lloyds Bank Rev.* 74 (October): 25–35. Also in [C.6].

———. "The Economics of Brain Drain." In *Sixth Seminar on Canadian-American Relations*, edited by F. J. Boland. Windsor, Ont.: Windsor Univ. Press. Also as "The Economics of the 'Brain Drain': The Canadian Case." *Minerva* 3 (Spring 1965): 299–311. Also as "The Technology Gap and the Brain Drain: III." In *Defense, Science, and Public Policy*, edited by Edwin S. Mansfield. New York: Norton, 1968.

———. "Employment Theory and Public Policy in the North American Context." In *Employment, Unemployment and Manpower*, edited by Frances Bairstow. Montreal: McGill Univ. Press (for Indus. Relations Centre).

———. "An Essay on Advertising." *Malayan Econ. Rev.* 9 (October): 6–9.

———. "Fiscal Policy and the Balance of Payments in a Growing Economy." *Malayan Econ. Rev.* 9 (April): 1–13. Also in *Government Finance and Economic Development*, edited by Alan T. Peacock and Gerald Hauser. Paris: OECD, 1965. Also in *Readings on Taxation in Developing Countries*, rev. ed., edited by Richard M. Bird and Oliver Oldman. Baltimore: Johns Hopkins Univ. Press, 1967. Also in [A.10].

———. "The International Competitive Position of the United States and the Balance of Payments Prospect for 1968." *Rev. Econ. and Statis.* 46 (February): 14–32.

———. "Johnson's Northern Tour: A Traveller's Guide Past the Marshes." *Canadian J. Econ. and Polit. Sci.* 30 (August): 435–38.

———. "Major Issues in Monetary and Fiscal Policies." *Federal Reserve Bull.* 50 (November): 1400–1413. Also in [A.10]. Also in *Monetary Economics: Readings*, edited by Alan D. Entine. Belmont, Calif.: Wadsworth, 1968. Also in *Readings in Money, National Income, and Stabilization Policy*, rev. ed., edited by Warren L. Smith and Ronald L. Teigen. Rev. ed. Homewood, Ill.: Irwin, 1970.

———. "The New Tariff Policy for the Automotive Industry." *Bus. Q.* 29 (Spring): 43–57.

———. "The Ottawa-Washington Troubles." *Challenge* 12 (June): 25–27. Also in *American Foreign Economic Policy: Essays and Comments*, edited by Benjamin J. Cohen. New York: Harper & Row, 1968.

———. "Statement, the Federal Reserve System after Fifty Years." Hearings before the Subcommittee on Domestic Finance of the Committee on Banking and Currency, House of Representatives, 88th Cong., 2d sess.

———. "Stefan Stykolt: A Personal Memoir." In *The Living Name: A Tribute to Stefan Stykolt from Some of His Friends*, edited by W. J. Stankiewicz. Toronto: Univ. Toronto Press.

———. "Tariffs and Economic Development: Some Theoretical Issues." *J. Development Studies* 1 (October): 3–30. Also in *Economics of Trade and Development,* edited by James D. Theberge. New York: Wiley, 1968. Also in [A.15].

1965. "Approaches to the Reduction of Poverty: II." *A.E.R. Papers and Proc.* 55 (May): 541–48. Also in *Problems of the Modern Economy,* edited by Edmund S. Phelps. New York: Norton, 1966. Also in *Inequality and Poverty,* edited by Edward C. Budd. New York: Norton, 1967.

———. "The Costs of Protection and Self-Sufficiency." *Q.J.E.* 79 (August): 356–72. Also in [A.15].

———. "The Decade Ahead: In U.S. International Economic Relations." *J. Bus.* 38 (July): 245–51. Also as "America in the World Economy—the Decade Ahead." Selected Paper no. 17, Grad. School Bus., Univ. Chicago, n.d.

———. "An Economic Theory of Protectionism, Tariff Bargaining, and the Formation of Customs Unions." *J.P.E.* 73 (June): 256–83. Also in [A.15].

———. "Economics and Politics of Opulence." *Univ. Toronto Q.* 34 (July): 313–31.

———. "Federal Support of Basic Research: Some Economic Issues." In *Basic Research and National Goals,* by the National Academy of Science. Report to the Committee on Science and Astronautics, U.S. House of Representatives. Washington: Government Printing Office. Also in *Minerva* 3 (Summer): 500–514. Also as "Paying for Basic Research." *Bull. Atomic Scientists* 21 (December): 12–16. Also in *Defense, Science, and Public Policy,* edited by Edwin S. Mansfield New York: Norton, 1968.

———. "International Monetary O
Money and Finance: Readings in Th
by Deane Carson. New York: Wil

———. "Is Inflation the Inevitable
Retarding Factor in Economic
L'Actualité écon. 41 (January/March
Econ. Rev. 11 (April 1966): 21–2
financieras 5 (March/April 1966) 369
Retarding Factor in Economic Growth?" In *Fiscal and*
Problems in Developing States, edited by David Krivine New York, Praeger, 1967. Also in [A.10]. Also as "Inflation and Development Policy." In *Leading Issues in Economic Development,* 2d ed., edited by Gerald M. Meier. New York: Oxford Univ. Press, 1970. 3d ed., 1976; 4th ed., 1984.

———. "Notes on the Theory of Economic Policy under Fixed Exchange Rates." *Osaka Econ. Papers* 14 (December): 45–48.

———. "Optimal Trade Intervention in the Presence of Domestic

Distortion." In *Trade, Growth and the Balance of Payments: Essays in Honor of Gottfried Haberler,* by Robert E. Baldwin et al. Amsterdam: North-Holland. Also in *International Trade: Selected Readings,* edited by Jagdish N. Bhagwati. Baltimore: Penguin, 1969; enl. ed., Cambridge, Mass.: M.I.T. Press, 1981. Also in [A.15].

———. "Poverty and Unemployment." In *The Economics of Poverty: An American Paradox,* edited by Burton A. Weisbrod. Englewood Cliffs, N.J.: Prentice-Hall.

———. "Productivity Trends in British University Education: Correspondence." *Minerva* 4 (Autumn): 95–98.

———. "A Quantity Theorist's Monetary History of the United States." *Econ. J.* 75 (June): 388–96.

———. "Roosa on International Monetary Reform." *Nat. Banking Rev.* 3 (December): 187–92.

———. "Should Gold Be Scrapped?" *Nat. Banking Rev.* 2 (June): 455–60. Also in *J. Liberal Thought* 1:7–14. Also in *Empire Club Addresses, 1964–65.* Also in *Money and Economic Activity: Readings in Money and Banking,* 3d rev. ed., edited by Lawrence S. Ritter. New York: Houghton Mifflin, 1967.

———. "Some Aspects of the Theory of Economic Policy in a World of Capital Mobility." *Riv. internaz. sci. econ. e commerciali* 12 (June): 545–59. Also in *Essays in Honour of Marco Fanno,* edited by Tullio Bagiotti. Padua: CEDAM, 1966. Also in [A.18].

———. "A Theoretical Model of Economic Nationalism in New and Developing States." *Polit. Sci. Q.* 80 (June): 169–85. Also in [C.6]. Also in German in *Zwei Ansätze der politischen Ökonomie,* edited by Bruno S. Frey and Werner Meissner. Frankfurt: Verlag, 1974.

———. "The Theory of Tariff Structure, with Special Reference to World Trade and Development." Reprint ser. no. 17, Comm. Comparative Study of New Nations, Univ. Chicago. Also in [A.7; A.15].

———. "U.S. Economic Policy toward the Less Developed Countries: A Survey of Major Issues." In *Measures for Trade Expansion of Developing Countries,* Research Paper no. 5. Tokyo: Japan Econ. Res. Centre. Also in Spanish in *Cuadernos econ.* 3 (August 1966): 1–17.

———. "United States Policy and the Problems of Developing Countries." *J. Bus.* 38 (October): 337–43. Also a summary in Italian in *Mercurio* 9 (June 1966): 8–12.

1966. "Balance-of-Payments Controls and Guidelines for Trade and Investment." In *Guidelines, Informal Controls, and the Market Place: Policy Choices in a Full Employment Economy,* edited by George P. Shultz and Robert Z. Aliber. Chicago: Univ. Chicago Press.

———. "Canadian-American Economic Integration: A Time for Decision." *J. Canadian Studies* 1 (August): 31–36.

———. "Canadian-American Economic Relations." In *Problems of Journalism: Proceedings of the American Society of Newspaper Editors.* New York: American Soc. Newspaper Editors.

———. "The Commonwealth Preferences: A System in Need of Analysis." *Round Table* 224 (October): 363–78.

———. "The Economics of Brain Drain: Correspondence." *Minerva* 4 (Winter): 273–74.

———. "Factor Market Distortions and the Shape of the Transformation Curve." *Econometrica* 34 (July): 696–98. Also in [A.15].

———. "International Economics: Progress and Transfer of Technical Knowledge: Discussion." *A.E.R. Papers and Proc.* 56 (May). 280–83.

———. "International Monetary Reform and the Less Developed Countries." *Malayan Econ. Rev.* 11 (April): 1–20. Also in [A.9; A.10].

———. "A Model of Protection and the Exchange Rate." *Rev. Econ. Studies* 33 (April): 159–63. Also in [A.15].

———. "The Neo-classical One-Sector Growth Model: A Geometrical Exposition and Extension to a Monetary Economy." *Economica* 33 (August): 265–87. Also in [A.10]. Also as "Money in a Neo-classical One-Sector Growth Model." In *Monetary Theory: Selected Readings*, edited by Robert W. Clower. Baltimore: Penguin, 1969. Also in *Macroeconomic Themes: Edited Readings in Macroeconomics*, edited by M. J. C. Surrey. London: Oxford Univ Press, 1976

———. "A Note on Tariff Valuation Bases, Economic Efficiency and the Effects of Preferences." *J P E* 74 (Aug

———. "Notes on the Welfare Cost of

Philippine Econ. J. 5 (2): 277–88.

———. "The Objectives of Economic

Monetary Policy under Fixed Excha

Restoring Balance in International Pa

Princeton, N.J.: Princeton Univ. Pr

naz. sci. econ. e commerciali 13 (June)

———. "Possibilities and Limitations of Joint Planning

American Planning: The Seventh Annual Conference on Canadian-American Relations, edited by F. J. Boland. Toronto. Univ Toronto Press (for Univ. Windsor).

———. "Préférences commerciales, accords régionaux et développement économique." *Rev. écon.* 6 (November): 896–931

———. "The Social Sciences in the Age of Opulence." *Canadian J Econ. and Polit. Sci.* 32 (November): 423–42. Also in [A.23].

———. "Sugar Protectionism and the Export Earning of Less Developed Countries: Variations on a Theme by R. H. Snape." *Economica* 33 (February): 34–42.

———. "Trade Preferences and Developing Countries." *Lloyds Bank Rev.* 80 (April): 1–18. Also in German in *Wirtschaftsdienst* 46 (July): 367–76. Also in Italian and English in *Cronache econ.*, nos. 299–300 (November/December 1967), pp. 29–47. Also in *Reshaping the World Economy: Rich and Poor Countries*, edited by John A. Pincus. Englewood Cliffs, N.J.: Prentice-Hall, 1968. Also in *Expansion of World Trade and the Growth of National Economies: Significant Papers*, edited by Richard S. Weckstein. New York: Harper & Row, 1968. Also in *Economic Development of Tropical Agriculture: Theory, Policy, Strategy, and Organization*, edited by W. W. McPherson. Gainesville: Univ. Florida Press, 1968; French edition, 1972. Also in *International Trade and Finance*, edited by Robert E. Baldwin and J. David Richardson. Boston: Little, Brown, 1974.

———. "Unemployment and Poverty." In *Poverty amid Affluence*, edited by Leo Fishman. New Haven, Conn.: Yale Univ. Press.

———. "The Welfare Costs of Exchange-Rate Stabilization." *J.P.E.* 74 (October): 512–18. Also in [A.18].

1967. "The Annual Meetings." *J.P.E.* 75 (December): 893.

———. "Are You Worth Your Weight in Gold?" *J.P.E.* 75 (April): 205–7. Also in [A.18].

———. "The Atlantic Case." *New Society* 9 (May 18): 724–27. Also in *Atlantic Community* 5 (Fall): 371–80.

———. "Comment on the de Leeuw Findings on Demand for Money." In *Monetary Process and Policy: A Symposium*, edited by George Horwich. Homewood, Ill.: Irwin.

———. "Economic Policy in a World of Capital Mobility: Some Corrections" (with James B. March). *Riv. internaz. sci. econ. e commerciali* 14 (November): 1127–31.

———. "Economic Relations between the United States, Europe and the Developing Countries: The Need for New Policies in the Spirit of the Marshall Plan." Paper presented at a Brussels conference on the twentieth anniversary of General Marshall's Harvard speech. *Punjab Univ. Econ.* 5 (June): 1–8.

———. "The Economics of Advertising: A Review of the EAG Report." *Advertising Q.* 13 (Autumn): 9–14.

———. "The Ideology of Economic Policies in the New State." Paper presented at a seminar of the Committee for Comparative Study of New Nations. In [C.6].

———. "The International Circulation of Human Capital: Comment." *Minerva* 6 (Autumn): 105–12. Also as "Criticisms of Thomas's Analysis of Brain Drain." In *Economics of Education: Selected Readings*, vol. 2, edited by Mark Blaug. Harmondsworth: Penguin, 1969.

———. "International Trade in Agricultural Products in Relation to

Economic Development: Comments." In *Agricultural Development and Economic Growth*, edited by Herman M. Southworth and Bruce F. Johnston. Ithaca, N.Y.: Cornell Univ. Press.

———. "International Trade Theory and Monopolistic Competition Theory." In *Monopolistic Competition Theory: Studies in Impact*, edited by Robert E. Kuenne. New York: Wiley.

———. "Issues in Monetary Research: Prefatory Note." *J.P.E.* 75, no 4, pt. 2 (August): 431–32.

———. "The Kennedy Round." *J. World Trade Law* 1 (July/August): 475–81. Also in *The World Today* 23 (August): 326–33.

———. "The Neutrality of Money in Growth Models: A Reply." *Economica* 34 (February): 73–74.

———. "New Methods of Price Measurement: Discussion." In *Proceedings of the Social Statistics Section's Annual Meeting*. Washington: American Statis. Assoc.

———. "Nominal Tariff Rates and United States Valuation Practices: Two Case Studies" (with Herbert G. Grubel). *Rev. Econ. and Statis.* 49 (May): 138–42.

———. "Nominal Tariffs, Indirect Taxes and Effective Rates of Protection: The Common Market Countries 1959" (with Herbert G. Grubel). *Econ. J.* 77 (December): 761–76.

———. "Policy Obstacles to Trade and Development." In [A.9]. Also in *Economics of Trade and Development*, edited by James D. Theberge. New York: Wiley, 1968.

———. "The Possibility of Factor Price Equalization When Commodities Outnumber Factors." *Econ J* [illegible]

———. "The Possibility of Income Lo[illegible]
or Factor Accumulation in the Pr[illegible]
(March): 151–54. Also in *Internat*[illegible]
edited by Jagdish N. Bhagwati. Ca[illegible]
1981.

———. "A Reply to Professor Dales." *J* [illegible]
36–39.

———. "The Report on Bank Charges [illegible]
(August): 64–68. Also in [C.18].

———. "Some Economic Aspects of Brain Drain" with appendix "Notes on the Effects of Migration of Professional People on the Welfare of Those Remaining Behind." *Pakistan Development Rev.* 7 (Autumn): 379–411. Also as "An 'Internationalist' Model." In *The Brain Drain*, edited by Walter Adams. New York: Macmillan, 1968. Also as "International Circulation of Human Capital or 'Brain Drain' " in [A.24].

———. "Theoretical Problems of the International Monetary System." *J. Econ. Studies* 2 (January): 3–33. Also in *Pakistan Development*

Rev. 7 (Spring): 1–28. Also in Italian in *Riv. polit. econ.* 58 (August/September 1968): 1111–45. Also in *International Finance: Selected Readings*, edited by Richard N. Cooper. Harmondsworth: Penguin, 1969. Also in [A.18].

———. "U.S. Trade Policies toward Less-developed Countries." In [A.9]. Also in *American Foreign Economic Policy: Essays and Comments*, edited by Benjamin J. Cohen. New York: Harper & Row, 1968.

———. "The Warrants for Basic Research: A Comment." *Minerva* 5 (Winter): 269–73.

1968. "Advantages, Disadvantages and Alternatives for Britain." In *Euroforum 68: Europe in the World of Tomorrow*, by the Economic Council of the CDU. Saarbrucken: Econ. Council CDU.

———. "Alternative Maximization Policies for Developing-Country Exports of Primary Products." *J.P.E.* 76 (May/June): 489–93.

———. "Balogh and Powell on Myrdal." *Bull. Inst. Development Studies* 1 (October): 20–24.

———. "Canadian Contributions to the Discipline of Economics since 1945." *Canadian J. Econ.* 1 (February): 129–46.

———. "A Catarrh of Economists? From Keynes to Postan." *Encounter* 30 (May): 50–54.

———. "Current Issues in Monetary Policy." *Financial Analysts J.* 24 (July/August): 1–6. Also in *Bankers Magazine* 206 (November): 251–57. Also in *Malaysian Management Rev.* 3 (December): 55–64. Also in *Money and Banking: Theory, Analysis, and Policy*, edited by S. Mittra. New York: Random House, 1970.

———. "Economic Theory and Contemporary Society." *Univ. Toronto Q.* 37 (July): 321–37. Also in *Economics* 8 (Spring/Summer 1969): 1–15. Also in [A.23].

———. "The Economics of Student Protest." *New Society* 12 (November 7): 673–75. Also in [C.20].

———. "Ely Devons: Obituary." *LSE Magazine* 35 (June): 13.

———. "The Gain from Exploiting Monopoly or Monopsony Power in International Trade." *Economica* 35 (May): 151–56. Also in [A.15].

———. "Harmonization of Economic Policies under Free Trading Arrangements: Issues for Canada." *J. Canadian Studies* 3 (November): 16–21.

———. "The Implications of Free or Freer Trade for the Harmonization of Other Policies." In *Economic Development of Tropical Agriculture: Theory, Policy, Strategy, and Organization*, edited by W. W. McPherson. Gainesville: Univ. Florida Press; French edition, 1972. Also in [A.13; A.15].

———. "International Trade: Theory." In *International Encyclopedia of the Social Sciences*, vol. 8, edited by David L. Sills. New York: Macmillan.

———. "Is There an American Threat?" In *The U.S. and Us*, edited by Gordon McCaffrey. Toronto: Canadian Inst. Public Affairs.

———. "LDC Investment: The Road Is Paved with Preferences." *Columbia J. World Bus.* 3 (January/February): 17–21.

———. "The Monetary Implications of Free Trade Association." In *A Free Trade Association*, edited by Thomas M. Franck and Edward Weisband. London: Weidenfeld & Nicolson.

———. "NAFTA: Monetary System—More of the Same." *Columbia J. World Bus.* 3 (September/October): 57–60.

———. "A Note on Seigniorage and the Social Saving from Substituting Credit for Commodity Money." *Univ. Punjab Econ.* 6 (June): 1–8. Also in *Monetary Problems of the International Economy*, edited by Robert A. Mundell and Alexander K. Swoboda. Chicago: Univ. Chicago Press, 1969.

———. "Notes on Some Theoretical Problems Posed by the Foreign Trade of Centrally Planned Economies." In *International Trade and Central Planning: An Analysis of Economic Interactions*, edited by Alan A. Brown and Egon Neuberger. Berkeley: Univ. California Press.

———. "Observations on the Bank Merger Proposals." *Bankers Magazine* 206 (September): 135–37.

———. "Priorities—a Question but Not an Answer." *Polit. Q.* 39 (October/December): 371–75.

———. "Problems of Balance-of-Payments Adjustment in the Modern World." In *Essays in Money and Banking in Honour of R. S. Sayers*, edited by C. R. Whittlesey and J. S. G. Wilson. Oxford: Clarendon

———. "Problems of Efficiency in Mone

(September/October): 971–90. Also

1 (2): 127–51. Also in [A.18; C.18]

———. "The Sterling Crisis of 1967 a

Woody Thompson Memorial Lect

Economic Association, April. *Neb*

tumn): 3–17.

———. "A Strategy for UNCTAD." *Round*

13.

———. "Towards a New Basis for the World's Money." *The Times* (March 18): 20.

———. "U.S. Economic Policy toward the Developing Countries: Summary of Proceedings." *Econ. Development and Cultural Change* 16 (April): 357–84.

———. "The Watkins Report: Towards a New National Policy?" *Internat. J.* 23 (Autumn): 615–22. Also in *Canadian Economic Issues*, edited by L. D. Pal. Toronto: Macmillan, 1971.

———. "Why a Wealth Tax?" *Spectator* 221 (December 27): 900–901.

———. "World Trade Policy in the Post-Kennedy Round Era: A Survey of Alternatives with Special Reference to the Position of the

Pacific and Asian Regions." *Econ. Record* 44 (June): 154–67. Also as "A New World Trade Policy in the Post–Kennedy Round Era." In *Pacific Trade and Development: Proceedings of a Conference,* edited by Kiyoshi Kojima. Tokyo: Japan Econ. Res. Centre.

1969. "The Case for Flexible Exchange Rates, 1969." In *U.K. and Floating Exchanges: A Debate on the Theoretical and Practical Implications,* by Harry G. Johnson and John E. Nash. Hobart paper no. 46. London: Inst. Econ. Affairs. Also in *Federal Reserve Bank of St. Louis Rev.* 51 (June): 12–24. Also in *Approaches to Greater Flexibility of Exchange Rates: The Burgenstock Papers,* edited by George N. Halm. Princeton, N.J.: Princeton Univ. Press, 1970. Also as "Should Foreign Exchange Rates Be Pegged—or Float?" In *Economics: A Reader,* edited by Kenneth G. Elzinga. New York and London: Harper & Row, 1972. Also in *Money and Finance: Readings in Theory, Policy, and Institutions,* 2d ed., edited by Deane Carson. New York: Wiley, 1972. Also in [A.18]. Also in *International Trade and Finance,* edited by Robert E. Baldwin and J. David Richardson. Boston: Little, Brown, 1974.

———. "The Case for a Floating Exchange Rate." *Spectator* 222 (April 18): 496–97.

———. "Current International Economic Policy Issues." *J. Bus.* 42 (January): 12–21.

———. "The Decline of the International Monetary System." *World Today* 25 (March): 103–9. Also in *Mercurio* 12 (August): 1–7.

———. "Excise Taxes and Effective Protection: A Note" (with Herbert G. Grubel and William V. Rapp). *Econ. J.* 79 (September): 674–75.

———. "Financial and Monetary Problems: Britain and the EEC." In *Economics: Britain and the EEC,* edited by M. A. G. van Meerhaege. London: Longmans. Also as "Financial and Monetary Problems of the United Kingdom." *J World Trade Law* 3 (August): 364–73. Also in *UMBC Econ. Rev.* 5 (2): 31–37.

———. "Further Stress Ahead for Monetary System." *Financial Times* (December 31), p. 12.

———. "The Future of Gold and the Dollar." *J. World Trade Law* 3 (March/April): 117–29.

———. "The Gold Rush of 1968 in Retrospect and Prospect." *A.E.R. Papers and Proc.* 59 (May): 344–48.

———. "Inside Money, Outside Money, Income, Wealth, and Welfare in Monetary Theory." *J. Money, Credit and Banking* 1 (February): 30–45. Also in [A.18].

———. "The International Monetary Crisis." Third Annual Monash University Economics Lecture. In *Studies in International Economics: Monash Conference Papers,* edited by I. A. McDougall and Richard

H. Snape. Amsterdam: North-Holland, 1970. As "The World Monetary Crisis: Gold, Dollars and the 'Crawling Peg.' " *Encounter* 35 (August 1970): 43–52. Also in [A.18].

———. "The International Monetary Problem: Gold, Dollars, Special Drawing Rights, Wider Bands and Crawling Pegs." University of Calgary Distinguished Lecture Series. Also in *Linking Reserve Creation and Development Assistance.* Hearings of the Subcommittee on International Trade and Payments of the Joint Economic Committee of Congress. 91st Cong., 2d sess., May 28. Washington: Government Printing Office.

———. "Minimum Wage Laws: A General Equilibrium Analysis." *Canadian J. Econ.* 2 (November): 599–604.

———. "A Monetary View of the International Monetary Problem." In *The U.K., U.S.A., and Europe in the World Economy,* by the Business Economists Group. London: Bus. Econ. Group.

———. "The North-South Problem in the World Economy and the Implications of UNCTAD 1968." In *Pacific Trade and Development II: Papers and Proceedings,* edited by Kiyoshi Kojima. Tokyo: Japan Econ. Res. Centre. Also in *Foreign Trade Rev.* 4 (April/June): 1–13. Also in [C.16].

———. "Notes on the Geometry of Income Distribution in a Two-Factor, Two-Commodity Model." *Osaka Econ. Papers* 17 (March): 31–38.

———. "On Demolishing the Barriers to Trade." In *Rebuilding the Liberal Order,* by the Institute for Economic Affairs. London: Inst. Econ. Affairs.

———. "Pesek and Saving's Theory of Money and Wealth: A Comment." *J. Money, Credit and Banking* 1 (Aug

———. "The 'Problems' Approach to

form." In *Monetary Problems of the I*

Robert A. Mundell and Alexander

Chicago Press. Also in [A.18].

———. "Reply by Professor Johnson t

In *International Economic Relations: Pr*

the International Economic Association, edited

New York: Macmillan.

———. "Reserve Creation and Development Assistance." In *Linking Reserve Creation and Development Assistance.* Hearings before the Subcommittee on International Exchange and Payments of the Joint Economic Committee of Congress. 91st Cong., 2d sess., May 28. Washington: Government Printing Office.

———. "Some Aspects of the Multilateral Free Trade Association Proposal." *Manchester School Econ. and Soc. Studies* 37 (September): 189–212.

———. "The Standard Theory of Tariffs." *Canadian J. Econ.* 2 (August): 333–52.

———. "The Theory of Effective Protection and Preferences." *Economica* 36 (May): 119–38. Also in [A.15].

———. "The Theory of International Trade." In *International Economic Relations: Proceedings of the Third Congress of the International Economic Association*, edited by Paul A. Samuelson. New York: Macmillan. Also in French in *L'Actualité écon.* 44 (January/March): 621–37.

1969–70. "Trade Challenges Confronting Commonwealth Countries." *Internat. J.* 25 (Winter): 109–28. Also in *Commonwealth Policy in a Global Context*, edited by Paul Streeten and Hugh Corbet. London: Cass, 1971.

1970. "Border Taxes, Border Tax Adjustments, Comparative Advantage, and the Balance of Payments" (with Melvyn B. Krauss). *Canadian J. Econ.* 3 (November): 595–602. Also in [A.18].

———. "Britain and the Six." Introduction to *U.K., Commonwealth, and Common Market: A Reappraisal*, 3d ed., by James E. Meade. London: Inst. Econ. Affairs.

———. "Canada's Floating Dollar in Historical Perspective." *Internat. Currency Rev.* 2 (July/August): 4–9.

———. "The Case for a Multilateral Free Trade Association: The Asian Interest." *Philippine Econ. J.* 9 (1): 64–82.

———. "The Challenge of Political, Economic and Social Forces for the Financial System." Paper presented at the First National Conference of Canadian Bankers. *ICS Bull.* 3 (Fall): 10–14. As "The Future of the Financial System." *Bankers Magazine* 211 (February 1971): 59–64. Also in *UMBC Econ. Rev.* 7, no. 1 (1971): 31–38.

———. "Comment on Cooper's Paper." In *International Reserves: Needs and Availability*, by the International Monetary Fund. Washington: Internat. Monetary Fund.

———. Comment on Corden's "The Efficiency Effects of Trade and Protection." In *Studies in International Economics: Monash Conference Papers*, edited by I. A. McDougall and Richard H. Snape. Amsterdam: North-Holland.

———. "Consumer's Surplus and Micro-Macro: Comment." *J.P.E.* 78 (January/February): 137–38.

———. "Developments in Monetary Theory since Radcliffe." *Bankers Magazine* 209 (January): 9–13.

———. "Direct Foreign Investment: Survey of the Issues." Paper presented at the Third Pacific Conference on Trade and Development, Sydney, Australia. In *Direct Foreign Investment in Asia and the Pacific*, edited by Peter Drysdale. Canberra: Australian National Univ. Press, 1972. Also in [A.23; A.24].

———. "Directives à l'intention des gouvernements au sujet des sociétés multinationales." *Rev. Societé d'Etudes et d'Expansion,* no. 243 (November/December), pp. 888–94. Also as pamphlet no. 367, CNBOS, Brussels, 1971.

———. "The Economics of Advertising." In *Readings in the Economics of Industrial Organization,* edited by Douglas Needham. New York: Holt, Rinehart, & Winston.

———. "The Effects of Unionization on the Distribution of Income: A General Equilibrium Approach" (with Peter Mieszkowski). *Q.J.E* 84 (November): 539–61.

———. "The Efficiency and Welfare Implications of the International Corporation." In *The International Corporation: A Symposium,* edited by Charles P. Kindleberger. Cambridge, Mass.: M.I.T. Press. Also in *Studies in International Economics: Monash Conference Papers,* edited by I. A. McDougall and Richard H. Snape. Amsterdam: North-Holland.

———. "Emigration and the Supply and Demand for Medical Manpower: The Irish Case." *Minerva* 8 (January): 144–46.

———. "Further Notes on the Geometry of Income Distribution." *Osaka Econ. Papers* 18 (March): 27–34.

———. "The Future of Gold, If Any: Demonetization vs. Remonetization" (L'Or a-t-il un avenir?). *Écon. appl.* 23 (1). 139–48

———. "Government and the Corporation: A Fallacious Analogy." In *Approaches to Greater Flexibility of Exchange Rates. The Burgenstock Papers,* edited by George N. Halm. Princet
Press.

———. "International Trade and Pa
23d International Banking Summe
ers.

———. "Is There an Optimal Mon
(May): 435–42. Also in *Frontiers of Q*
Michael D. Intriligator. Amsterdam. Nor
[A.18].

———. "Keynes & the Keynesians: Some Intellectual Legends." *Encounter* 34 (January): 70–73. Also in [A.18]

———. "Labour Mobility and the Brain Drain." Paper presented at the International Economic Association Conference, Bled, Yugoslavia. In *The Gap between Rich and Poor Nations,* edited by Gustav Ranis. London: Macmillan, 1972.

———. "The Manpower Dimensions of Economic Growth. Discussion." In *Proceedings* of the 1969 Annual Winter Meeting, Industrial Relations Research Assoc., edited by Gerald G. Somers. Madison, Wis.: Industrial Relations Res. Assoc.

———. "Memorandum . . . on Overseas Aid." In *Minutes of Evidence,* by the Select Committee on Overseas Aid, Subcommittee A. London: H.M.S.O.

———. "Monetary Theory and Monetary Policy." *Euromoney* 2 (December): 16–20. Also in [A.18].

———. "The Multinational Corporation as an Agency of Economic Development: Some Exploratory Observations." In *The Widening Gap: Development in the 1970's*, edited by Barbara A. Ward, Leonid Anjou, and J. D. Runnallas. New York: Columbia Univ. Press.

———. "A New View of the Infant Industry Argument." In *Studies in International Economics: Monash Conference Papers*, edited by I. A. McDougall and Richard H. Snape. Amsterdam: North-Holland.

———. "North American Policies for the World beyond 1967." In *Canada and the United States in the World of the Seventies*, proceedings of the Ninth Annual Conference on Canadian-American Relations, edited by R. H. Wagenburg. Windsor, Ont.: Univ. Windsor Press.

———. "A Note on Distortions and the Rate of Growth of an Open Economy." *Econ. J.* 80 (December): 990–92.

———. "A Note on the Theory of Transactions Demand for Cash." *J. Money, Credit and Banking* 2 (August): 383–84.

———. "Objectives and Strategies for Development." In *U.S. Policies toward Developing Countries* Hearings before the Subcommittee on Foreign Economic Policy of the Joint Economic Committee, 91st Cong., 2d sess. Washington: Government Printing Office. Also in *UMBC Econ. Rev.* 6 (2): 23–26.

———. "On Factor Price Equalization When Commodities Outnumber Factors: A Comment." *Economica* 37 (February): 89–90.

——— "Pacific Trade in an Open World" (with Hugh Corbet). *Pacific Community* 1 (April): 523–35.

———. "Pearson's 'Grande Assize' Fails: A Bleak Future for Foreign Aid." *Round Table* 60 (January): 17–25.

———. "Radcliffe—Ten Years On." *Euromoney* 1 (January): 12–13.

———. "Recent Developments in Monetary Theory: A Commentary." In [C.12]. Also in *UMBC Econ. Rev.* 7 (1): 16–22. Also in [A.18]. Also in German in *Geldtheorie*, edited by Karl Brunner, Hans G. Monissen, and Manfred J. M. Neumann. Cologne: Kiepenheuer & Witsch, 1974. Also in *Monetary Theory and Policy*, edited by Richard S. Thorn. New York: Praeger, 1976.

———. "Roy Harrod on the Price of Gold." In *Induction, Growth and Trade: Essays in Honour of Sir Roy Harrod*, edited by W. A. Eltis, M. F. G. Scott, and J. N. Wolfe. Oxford: Clarendon.

———. "The State of Theory in Relation to the Empirical Analysis." In *The Technology Factor in International Trade*, edited by Raymond Vernon. New York: Columbia Univ. Press (for N.B.E.R.).

———. "Summary of the Discussion." Paper presented at the Third Pacific Conference on Trade and Development, Sydney, Australia.

In *Direct Foreign Investment in Asia and the Pacific*, edited by Peter Drysdale. Canberra: Australian National Univ. Press, 1972.

———. "A Technical Note on the Width of the Band Required to Accommodate Parity Changes of Particular Size." In *Approaches to Greater Flexibility of Exchange Rates: The Burgenstock Papers*, edited by George N. Halm. Princeton, N.J.: Princeton Univ. Press.

———. "Thrust and Response: The Multinational Corporation as a Development Agent." *Columbia J. World Bus.* 5 (May/June): 25–30. Also in [A.24].

———. "Uniform versus Differentiated Protective Tariffs: A Comment." *J. Development Studies* 7 (October): 74–75.

———. "Verso una banca centrale mondiale: Intervento" [Toward a world central bank: discussion, with English summary]. *Bancaria* 26 (December): 1436–39.

———. ["The World Economy at the Crossroads"] (in Japanese). *Oriental Econ.*, no. 3519 (February 14), pp. 12–47.

1971. "Broad Trade Strategy for the Seventies." In [C.16].

———. "Commentary." On a paper by William M. Martin. In *Toward a World Central Bank?* Washington: Per Jacobsson Found.

———. "Comment on 'Macroeconomic Policy in an Open Economy' by Pascal Salin." Presented at the Conference on Stabilization Policies in Interdependent Economies, Paris, March. In *Stabilization Policies in Interdependent Economies*, edited by Emil Claassen and Pascal Salin. Amsterdam: North-Holland, 1972; in French, Paris: Editions Cujas, 1976.

———. "Comment on 'The Optimal Balance of Payments Deficit' by
Robert A. Mundell." Presented at the
Policies in Interdependent Econom
Policies in Interdependent Economies, e
cal Salin. Amsterdam: North-Hollan
tions Cujas, 1976.

———. "Comments on General Equili
tory." Presented at the Econometric

———. "Comments on Senator Grosart's Pa
ber): 544–47.

———. "Commercial Policy and Industrialization." Panama Lecture, August. *Economica* 39 (August 1972): 264–75. Also in [A.23].

———. "Common Market: The Case Against." *Spectator* 226 (February 13): 218.

———. "Concluding Remarks." Presented at the Conference on Stabilization Policies in Interdependent Economies, Paris, March. In *Stabilization Policies in Interdependent Economies*, edited by Emil Claassen and Pascal Salin. Amsterdam: North-Holland, 1972; in French, Paris: Editions Cujas, 1976.

———. "Controls versus Competition in Economic Development."

Lecture presented at the University of Ghana. *UMBC Econ. Rev.* 7 (2): 30–37. As "A Word to the Third World: A Western Economist's Frank Advice." *Encounter* 37 (October): 3–10.

———. "Crisis '71: On the Cards since 1958." *Money Management* (November/December), pp. 16–18. Also in [A.18].

———. "The Economic Benefits of Basic Research." *Minerva* 9 (April): 291–93.

———. "Effective Protection and General Equilibrium Theory." In [A.15]. Also in Spanish in *Información comercial española*, no. 463 (March 1972), pp. 3–13.

———. "Foreword." In *Essays in International Economics*, by J. Marcus Fleming. London: Allen & Unwin.

———. "Going into Europe—Again?" *Encounter* 36 (June): 9–10.

———. "Harking Back to Radcliffe." *Bankers Magazine* 212 (September): 115–20.

———. "Implications for the World Economy." In *Destiny or Delusion? Britain and the Common Market*, edited by Douglas Evans. London: Gollancz.

———. "The Independent University: Methods of Finance." In *University Independence: The Main Question*, by the Institute of Economic Affairs. London: Inst. Econ. Affairs.

———. "L'Inflation mondiale" [World inflation] (with Fritz Machlup, Assar Lindbeck, and Stephen Marris). *Banque* 302 (December): 155–70. Also in *Stabilization Policies in Interdependent Economies*, edited by Emil Claassen and Pascal Salin. Amsterdam: North-Holland, 1972; in French, Paris: Editions Cujas, 1976.

———. "Inflation: A Monetarist View." Discussion paper no. 1. London: Rowe, Rudd, & Co. Also in *Economic Policies in the 1970's*, edited by Alfred K. Ho. Ann Arbor: Grad. School Bus. Admin., Univ. Michigan. Also as "Inflation and World Trade: A 'Monetarist' View." *J. World Trade Law* 6 (January/February 1972): 9–19. Also in [A.18; A.23].

———. "The Inflation Crisis." *Internat. Currency Rev.* 3 (August): 5–11.

———. "International Economic Policy Problems." *Wall Street Transcript* 32 (June 21): 24.522–24.523.

———. "The Keynesian Revolution and the Monetarist Counter-revolution." *A.E.R. Papers and Proc.* 61 (May): 1–14. As "Revolution and Counter-revolution in Economics: From Lord Keynes to Milton Friedman." *Encounter* 36 (April): 23–33. Also in *CFA Digest* 2 (Summer 1972): 23–25. Also in [A.18]. Also in German in *Der Neue Monetarismus*. Nymphenburger Verlagstradlung, 1973. Also in [A.23; A.27].

———. "Notes on the Pattern of Interest Rates." *Osaka Econ. Papers* 20 (September): 17–29.

———. "Problems of European Monetary Union." *Euromoney* 2 (April): 39–43. Also in *J. World Trade Law* 5 (July/August): 377–87. Also in [A.18]. Also in *The Economics of Integration: A Book of Readings,* edited by Melvyn B. Krauss. London: Allen & Unwin, 1973. Also in *International Trade and Finance,* edited by Robert E. Baldwin and J. David Richardson. Boston: Little, Brown, 1974.

———. "Reflections on Current Trends in Economics." *Australian Econ. Papers* 10 (June): 1–11.

———. "Self-Assessment to Personal Income Tax: The American System." *British Tax Rev.*, no. 2 (March/April), pp. 78–85.

———. "Social, Political, and Economic Aspects of Development." Panama Lecture, August. In [A.23].

———. "The Special Economic Relationship between the United States and Canada." In *United States International Economic Policy in an Interdependent World,* vol. 2, by the Commission on International Trade and Investment Policy. Washington: Comm. Internat. Trade and Investment Policy.

———. "Testimony before the Standing Senate Committee on National Finance." In Ninth Proceedings on the Question of Growth, Employment and Price Stability, Senate of Canada, May 26. As "Inflation: An International Problem." *Bankers Magazine* 212 (July): 18–20.

———. "Theoretical Foundations of Monetary Policy: Discussion." In *Monetary Theory and Monetary Policy in the 1970s,* edited by G. Clayton, J. C. Gilbert, and R. Sedgwick. New York and London: Oxford Univ. Press.

———. "The Theory of Trade and G

sis." In *Trade, Balance of Payments, an*

Economics in Honor of Charles P. K

Bhagwati et al. Amsterdam: North

———. "Trade and Growth: A Geo

Econ. 1 (January/March): 83–101.

———. "The United States and the Internati

presented at the National War College, Washington, August

———. "Why Europe Must Sue U.S. for Trade Peace." *Guardian* (November 25), p. 23.

1972. "The Achievement of P. T. Bauer. Debunker of Economic Myths." *Encounter* 39 (November): 64–69.

———. "The Alternatives before Us." *J.P.E.* 80, no. 3, pt 2 (May/June): S280–S289.

———. "The Bretton Woods System, Key Currencies, and the 'Dollar Crisis' of 1971." *Three Banks Rev.*, no. 94 (June), pp. 3–22. As "The Bretton Woods System, Key Currencies, and the Rent Dollar Crises." In *Readings in Money, National Income, and Stabilization Pol-*

icy, 3d ed., edited by Warren L. Smith and Ronald L. Teigen. Homewood, Ill.: Irwin, 1974.

———. "Commercial Policy and the Monetary Crisis of 1971." In [C.19].

———. "The Common Market Questionnaire, October 1971" (with Mohsin S. Khan). *Economica* 39 (August): 316–22.

———. "Concluding Observations." Presented at the Wingspread Conference on National Monetary Policies and the International System, Racine, Wis. In *National Monetary Policies and the International Financial System,* edited by Robert Z. Aliber. Chicago: Univ. Chicago Press, 1974.

———. "The 'Crisis of Aid' and the Pearson Report." In *Foreign Resources and Economic Development: A Symposium on the Report of the Pearson Commission,* edited by T. J. Byres. London: Cass. Also in *International Economics and Development: Essays in Honor of Raul Prebisch,* edited by Luis Eugenio Di Marco. New York and London: Academic Press.

———. "The Dilution of Academic Power in Canada." *Minerva* 10 (July): 486–90.

———. "The Early Economics of Keynes." *Osaka Econ. Papers* 20 (March): 9–22. Also in *A.E.R. Papers and Proc.* 62 (May): 416–21.

———. "The Effect of the Monetary Problem on Development Cooperation: Linking Special Drawing Rights and Development." In *UNCTAD III Symposium.* The Hague: Nationale Commissie Ontwikkelingsstrategie 1970–80. Also in *Malayan Econ. Rev.* 17 (October): 1–10. Also in German in *Wirtschaftswoche* 36 (April 21): 37–41. Also as "The Link That Chains." *Foreign Policy,* no. 8 (Fall), pp. 113–20.

———. "A General Commentary." In *Bretton Woods Revisited,* edited by Keith Acheson, J. F. Chant, and M. F. J. Prachowny. Toronto: Univ. Toronto Press.

———. "General Statement: Proposals for Future Trade Strategy" (with Frank McFadzean et al.). In [C.19].

———. "Individual and Collective Choice." In *Man and the Social Sciences,* edited by William A. Robson. London: Allen & Unwin; Beverly Hills, Calif.: Sage, 1974.

———. "An Informal Classical Model of the Current Economic Development Problem." In *Economic Freedom, Stability and Growth,* edited by Willy Sellekaerts. Current Issues in Bus. and Econ., no. 1. Kalamazoo: Western Michigan Univ. Also in *Economic Development and Planning: Essays in Honor of Jan Tinbergen,* edited by Willy Sellekaerts. London: Macmillan, 1974. Also in *Portfolio: Internat. Econ. Perspectives,* vol. 3, no. 3 (1976).

———. "The International Monetary Crisis of 1971." Gilbert Lecture, Univ. Rochester. *Malayan Management Rev.* 7 (1): 41–52. Also in [A.18]. Also in *J. Bus.* 46 (January 1973): 11–23.

———. "The International Monetary System and the Rule of Law." *J. Law and Econ.* 15 (October): 277–92.

———. "The International Payments System." In *The World Banking Challenge: Lectures and Proceedings.* Twenty-fifth International Banking Summer School. Washington: American Bankers Assoc.

———. "Issues in Incomes Policy: Discussion." In *Proceedings* of the Twenty-fourth Annual Meeting of the Industrial Relations Research Association, edited by Gerald G. Somers. Madison, Wis.: Indus. Relations Res. Assoc.

———. "The Monetary Approach to Balance-of-Payments Theory." *J. Financial and Quantitative Analysis* 7 (March): 1555–72. Also in *Monte dei Paschi di Siena: Econ. Notes* 1 (January/April): 20–39. Also in [A.18]. Also in *International Trade and Money: The Geneva Essays,* edited by Michael B. Connolly and Alexander K. Swoboda. London: Allen & Unwin, 1973. Also in *Current Issues in Monetary Theory and Policy,* edited by Thomas M. Havrilesky and John T. Boorman. Arlington Heights, Ill.: AHM, 1976. Also in [C.24].

———. "The 1971 Currency Realignments and the Less Developed Countries: A Note on Revaluation and Inflation." *Philippine Econ. J.* 11 (1): 81–87.

———. "A Note on Dumping and Social Welfare" (with R. A. Cocks). *Canadian J. Econ.* 5 (February): 137–40.

———. "Notes on Incomes Policy and the Balance of Payments." In *Incomes Policy and Inflation,* edited by Michael P [illegible] Sumner. Manchester: Manchester U [illegible] *Note econ.* 2:5–17.

———. "Notes on the Economic Th [illegible] *Econ. Rev.* 17 (April): 1–7. Revised i [illegible] *tional Trade: Theory and Measurement,* [illegible] Amsterdam and Oxford: North-H [illegible] Elsevier, 1974.

———. "Optimal Negotiating Technique [illegible] Hugh Corbet). In [C.19].

———. "The Panamanian Monetary System." *Euromoney* 3 (January) 48–52. Also in [A.18].

———. "Political Economy Aspects of International Monetary Reform." *J. Internat. Econ.* 2 (September): 401–24. Also in [A.23].

———. "Problems of Stabilization Policy in an Integrated World Economy." In *Demand Management: Globalsteuerung. Symposium 1971,* edited by Herbert Giersch. Tubingen: Mohr. Also in [A.18]. Also as "The Theory of the Balance of Payments." In *International Trade and Finance,* edited by Robert E. Baldwin and J. David Richardson. Boston: Little, Brown, 1974.

———. "Some Economic Aspects of Science." *Minerva* 10 (January): 10–18. Also in *Civilization and Science, in Conflict or Collaboration?* A CIBA Foundation Symposium. Amsterdam: North-Holland.

———. "The Theory of Tax Incidence: A Diagrammatic Analysis" (with Melvyn B. Krauss). *Economica* 39 (November): 357–82.

———. "Trade and Growth: A Correction." *J. Internat. Econ.* 2 (February): 87–88.

———. "Trade, Investment and Labor, and the Changing International Division of Production." Paper presented at International Bank for Reconstruction and Development Conference on Industrialization and Trade Policies for the 1970s, October. In *Prospects for Partnership: Industrialization and Trade Policies in the 1970's*, edited by Helen Hughes. Baltimore: Johns Hopkins Univ. Press, 1973.

———. "Uncertainty and Probability in International Economics." In *Uncertainty and Expectations in Economics: Essays in Honor of G. L. S. Shackle*, edited by C. F. Carter and J. L. Ford. Oxford: Blackwell.

———. "We're Going to Have to Live with Bretton Woods." *Money Management* (March/April), pp. 30, 31, 56. Also in Spanish in *Rev. española de econ.* 2 (May/August): 84–103. Also in German in *Internationales Wahrunssystem und Inflation*, edited by Friedrich A. Lutz. Zurich: Schulthess, 1973.

———. "World Trading and Monetary Arrangements." In *Economics and World Order: From the 1970's to the 1990's*, edited by Jagdish N. Bhagwati. New York: Macmillan.

1973. "The Advantages of Integration." Comment on Mitchell Sharp's "Canada-U.S. Relations—Options for the Future." *Internat. Perspectives* (January/February), pp. 9–12.

———. "Economic Benefits of the Multinational Enterprise." In *Nationalism and the Multinational Enterprise: Legal, Economic, and Managerial Aspects*, edited by H. R. Hahlo, J. Graham Smith, and Richard W. Wright. Leiden: Sijthoff; Dobbs Ferry, N.Y.: Oceana.

———. "The Exchange-Rate Question for a United Europe: Internal Flexibility and External Rigidity versus External Flexibility and Internal Rigidity." In *Europe and the Evolution of the International Monetary System*, edited by Alexander K. Swoboda. Leiden: Sijthoff (for Grad. Inst. Internat. Studies).

———. "General Equilibrium with Public Goods: A Geometrical Analysis." In *Studies in Memory of Antonio de Viti de Marco*, vol. 2, edited by Ernesto d'Albergo. Bari: Cacucci.

———. "The Glass Bead Game of Orthodox Economics: Comment." *Intermountain Econ. Rev.* 4 (Spring): 72–73.

———. "The Glass Bead Game of Orthodox Economics: A Rejoinder." *Intermountain Econ. Rev.* 4 (Fall): 66.

———. "International Trade and Economic Development." In *Trade*

Strategies for Development, edited by Paul P. Streeten. London: Macmillan.

———. "Keynes, the Wider Band, and the Crawling Peg: A Comment" (with Elizabeth S. Johnson). *J. Money, Credit and Banking* 5 (November): 988–89.

———. "Narrowing the Exchange Rate Bands." In *European Monetary Unification and Its Meaning for the United States*, edited by Lawrence B. Krause and Walter S. Salant. Washington: Brookings Inst.

———. "National Styles in Economic Research: The United States, the United Kingdom, Canada, and Various European Countries." *Daedalus* 12 (Spring): 65–74. As "The Quantity and Quality of U.S. Economic Research." *Econ. Impact*, no. 6, pp. 46–50. Also in [A.23].

———. "Neglected Principles in the Negotiations on World Monetary Reform." Staff Paper no. 4. London: Trade Policy Res. Centre. As "The Optimal Solution for the Adjustment Problem." In *Zur Neuordnung des Internationalen Wahrunssystems*. Berlin: Duncker & Humblot. Also in *J. Econ. Studies*, N.S. 1 (May 1974): 17–29. Also as "General Principles for World Monetary Reform." In *Search of a New World Economic Order*, edited by Hugh Corbet and Robert Jackson. London: Croom Helm; New York: Wiley, 1974.

———. "A Note on Proportionally Distributed Quotas" (with Rachel McCulloch). *A.E.R.* 63 (September): 726–32.

———. "Notes on the Welfare Effects of a Reversed Transfer." *Osaka Econ. Papers* 21 (March): 45–52.

———. "On the Shape and Location of the Production Possibility Curve" (with Melvyn B. Krauss and Thanos Skouras) *Economica* 40 (August): 305–10.

———. "Reply to Mr. Michalopoulos [illegible] 43.

———. "Secular Inflation and the Int[illegible] *Money, Credit and Banking* 5, no. 1, [illegible]

———. "Some Micro-economic Refle[illegible] Inequalities." *Annals American Aca*[illegible] tember): 53–60. Also in [A.23].

———. "Summary of the Conference." In *European M*[illegible]*tion and Its Meaning for the United States*, edited by Lawrence B Krause and Walter S. Salant. Washington: Brookings Inst.

———. "The Uneasy Case for Universal Graduate Programmes in Economics." *Minerva* 11 (April): 263–68.

———. "The United States and the Disunited States. The Crisis of the International Economic System." In *Britain and the EEC*, edited by Douglas Evans. London: Gollancz.

———. "The University and Social Welfare: A Taxonomic Exercise." *Minerva* 11 (January): 30–52. Also in *Efficiency in Universities The*

La Paz Papers, edited by Keith G. Lumsden. Amsterdam: Elsevier, 1974. Also in [A.23].

1974. "The Chicago 'Core' Examination in Monetary Theory, Summer 1973." *Eastern Econ. J.* 1 (April): 111–24.

———. Commentary on C. B. MacPherson's "After Strange Gods: Canadian Political Science, 1973." In *Perspectives on the Social Sciences in Canada*, edited by Thomas N. Guinsburg and Grant Reuber. Toronto: Univ. Toronto Press.

———. "Comment on Lebedev." Presented at the Sixth Pacific Conference on Trade and Development, Mexico City.

———. Comment on "A Supranational Approach to the Analysis of Worldwide Inflation" by Walter S. Salant. Presented at the Brookings Conference on World Inflation, November. In *Worldwide Inflation: Theory and Recent Experience*, edited by Lawrence B. Krause and Walter S. Salant. Washington: Brookings Inst., 1977.

———. "The Current and Prospective State of Economics in Canada." In *Perspectives on the Social Sciences in Canada*, edited by Thomas N. Guinsburg and Grant Reuber. Toronto: Univ. Toronto Press. Also in *Australian Econ. Papers* 1 (June): 1–27. Also in [A.23].

———. "Economic Growth and Welfare: Correspondence." Comment on Beckerman's "Economic and Social Costs of Modernization and Development." *Minerva* 12 (January): 115–16.

———. "The Future of Floating Rates." Presented at a panel, Université de Paris-Dauphine, March. In French in *Banque* 20 (November): 1024–26. Also in *UMBC Econ. Rev.* 11, no. 1 (1975): 19–21. Also Part 2 of "Current Problems of the International Monetary System: Three Analyses" (with Rudiger Dornbusch and Jacob A. Frenkel). *Weltwirtschaftliches Archiv* 111, no. 2 (1975): 205–9. Also in *Recent Issues in International Monetary Economics*, vol. 2, edited by Emil Claassen and Pascal Salin Amsterdam: North-Holland, 1976.

———. "Gold, Money and the Law: Comment." Paper presented at a Conference on Gold-Clause Cases. In *Gold, Money and the Law*, edited by Henry G. Manne and Roger L. Miller. Chicago: Aldine, 1975.

———. "Introduction: Lessons of EFTA Experience in a Global Context." In *The Essentials of Economic Integration: Lesson of EFTA Experience*, by Victoria Curzon. London: Macmillan (for Trade Policy Res Centre).

———. "Issues in International Monetary Reform." *Portfolio: Internat. Econ. Perspectives*, vol. 2 (December).

———. "Major Issues in Monetary Economics." *Oxford Econ. Papers* 26 (July): 212–25. Also in Spanish in *Economía monetaria: Selección de textos*, edited by Mario I. Blejer and Arturo C. Porzecanski. Mexico City: Centro de Estudios Monetarios Latinoamericanos, 1977.

———. "Man and His Environment." *Osaka Econ. Papers* 22 (March): 33–64.

———. "Mercantilism: Past, Present and Future." *J. World Trade Law* 8 (January/February): 1–16. Also in *Manchester School Econ. and Soc. Studies* 42 (March): 1–17. Also in [C.23; A.23].

———. "The Monetary Theory of Balance-of-Payments Policies." *Nebraska J. Econ. and Bus.* 13 (Autumn): 144–63. Also in *Rev. española de econ.* 4 (May/August): 9–32. Also in [C.24].

———. "A Note on Optimum Distortion in International Trade: Reply." *J. Development Studies* 10 (January): 248.

———. "Notes on the Welfare Costs of Real and Money Transfers." *Eastern Econ. J.* 1 (July): 177–84.

———. "Observations on the Role of the University in Development Planning" (comment on a paper by D. Lim). *Minerva* 12 (January): 32–38.

———. "The Social and Intellectual Origins of *The General Theory*" (with Elizabeth S. Johnson). *Hist. Polit. Econ.* 6 (Fall): 261–77. Also in [A.27].

———. ["Some Reflections on Quesnay's *Tableau economique*"] (in Spanish). *Economía*, no. 2, pp. 76--85. Also in French in *Rev. econ. polit.* 85 (May/June 1975): 397–407.

———. "The State of Theory." *A.E.R. Papers and Proc.* 64 (May): 323–24.

———. "Technological Change and Comparative Advantage: An Advanced Country's Viewpoint." Paper [illegible] Sixth Pacific Conference on Trade and D
World Trade Law 9 (January/Februar
in *Información comercial española*, no
77.

———. "The Terms in Which Transf
Econ. J. 84 (March). 171.

———. "Trade-diverting Customs Unio
(September): 618–21.

———. "The Welfare Economics of Reversed International Transfers." In *Trade, Stability, and Macroeconomics. Essays in Honor of Lloyd A. Metzler*, edited by George Horwich and Paul A. Samuelson. New York and London: Academic Press.

———. "The Western Model of Economic Development." *World Development* 2 (February): 9–10.

———. "Why Devaluations Often Appear to Fail." *Eastern Econ. J.* 1 (October): 231–38. Also in Spanish in *Información comercial española*, no. 494 (October), pp. 51–59. Also in Italian in *Monte dei Paschi di Siena: Econ. Notes* 7 (January/February 1975): 5–18

1975. "Balance-of-Payments Adjustment with Bonds and Goods. A

Geometrical Approach." In *The Monetary Approach to the Balance of Payments.* London: Soc. Sci. Res. Council, Internat. Monetary Res. Programme. Also in *Pakistan Econ. and Soc. Rev.* 15 (Autumn/Winter 1977): 95–126.

———. "Chairman's Notes on the Symposium in Macroeconomic Models and Policy, Econometric Society World Congress, Toronto, Aug. 22, 1975." In *Frontiers of Quantitative Economics,* vol. 3, edited by Michael D. Intriligator. Amsterdam: North-Holland.

———. "The Classical Transfer Problem: An Alternative Formulation." *Economica* 42 (February): 20–31.

———. Commentary III on Herbert Giersch's "Freer Trade for Higher Employment and Price Level Stability." In *Toward a New World Trade Policy: The Maidenhead Papers,* edited by C. Fred Bergsten. Lexington, Mass.: Heath.

———. "Commercial Policy—What Is There Left? A Dialogue between Harry G. Johnson and a Government Official" (with Martin Lam). *J. World Trade Law* 9 (May/June): 341–47.

———. "The Denigration of Floating Exchange Rates." *Euromoney* (September), pp. 48–50.

———. "Egregious Economics as Pacifist Propaganda: Melman's Mythology of Defense Disbursements." *Armed Forces and Society* 1 (Summer): 498–504.

———. "Equity and Economic Theory." C. Woody Thompson Memorial Lecture, MEA. *Nebraska J. Econ. and Bus.* 14 (Summer): 3–17. As "Equity, Economic Theory, and Economic Development." *Portfolio: Internat. Econ. Perspectives,* vol. 3, no. 3 (1976). Also in *Against Equality: Readings on Economic and Social Policy,* edited by William Letwin. London: Macmillan, 1983.

———. "Foreword." In *Trade Effects of Public Subsidies to Private Enterprise,* by Geoffrey Denton, Seamus O'Cleireacain, and Sally Ash. London: Macmillan (for Trade Policy Res. Centre).

———. "A Formal Analysis of Some Brinley Thomas Problems Concerning the International Migration of Capital and Labour." *Osaka Econ. Papers* 25 (June): 24–36. Also in *Population, Factor Movements and Economic Development: Studies Presented to Brinley Thomas,* edited by Hamish Richards. Cardiff: Univ. Wales Press, 1976.

———. "Inflation, Energy and the World Economy: Prospects for International Cooperation." In *Selected Papers on Inflation/Recession, Energy and the International Financial Structure,* edited by Penelope Hartland-Thunberg. Washington: Georgetown Univ., Center Strategic and Internat. Studies.

———. "Inflation, Unemployment, and the Floating Rate." *Canadian Public Policy—Analyse polit.* 1 (Spring): 176–84.

———. "James Edward Meade: A Partial Tribute" (with A. Robert

Nobay). *Manchester School Econ. and Soc. Studies* 43 (September): 213–19.

———. "Keynes and British Economics." In *Essays on John Maynard Keynes,* edited by Milo Keynes. London and New York: Cambridge Univ. Press. Also in [A.23; A.27].

———. "Living with Inflation." *Banker* 125 (August): 863–64.

———. "The Monetary Approach to Balance-of-Payments Theory: A Diagrammatic Analysis." In *The Monetary Approach to the Balance of Payments.* London: Soc. Sci. Res. Council, Internat. Monetary Res. Programme. Also in *Manchester School Econ. and Soc. Studies* 43 (September): 220–74.

———. "Money and the Balance of Payments." University of Alabama Distinguished Lecture. In Spanish in *Información comercial española,* no. 506 (October), pp. 33–50. Also in Spanish in *CEMLA boletín mensual* 22 (January 1976): 10–17. Also in *Banca Naz. Lavoro Q. Rev.*, no. 116 (March 1976), pp. 3–18.

———. "A Note on Welfare-increasing Trade Diversion." *Canadian J. Econ.* 8 (February): 117–23.

———. "On Living without an International Monetary System." *Euromoney* (April), pp. 34–35.

———. "On the Record—Britain, Europe, and the World Economy." Address to the Foreign Affairs Club and Trade Policy Research Centre. *Banker* 125 (September): 48–50.

———. "The Organization and Retrieval of Economic Knowledge." Paper presented at an Internat. Econ. Assoc. Symposium, Kiel, Germany. As "Learning and Libraries. Academic Economics as a
Profession: Its Bearing on the Organisati
nomic Knowledge." *Minerva* 13 (Wi
gies of Economics." In *The Organi*
Knowledge, edited by Mark Perlman
Colo.: Westview, 1977.

———. "An Overview of the World
Kredit und Kapital 8 (4): 433–49. Als
economia 30 (January/March 1976): 5–2
Soc. Rev. 15 (Spring/Summer 1977). 1–16.

———. "The Probable Effects of Freer Trade on Individual Countries." In *Toward a New World Trade Policy: The Maidenhead Papers,* edited by C. Fred Bergsten. Lexington, Mass.: Heath

———. "Quadrangular Perspectives on Security, Inflation, and the World Financial Structure." In *Commissioned Papers on Inflation/Recession, Energy and the International Financial Structure,* edited by Penelope Hartland-Thunberg. Washington: Georgetown Univ., Center Strategic and Internat. Studies. As "Higher Oil Prices and the International Monetary System." *UMBC Econ. Rev.*, vol. 11, no.

1. Also in *The Economics of the Oil Crisis,* edited by T. M. Rybczynski. London: Macmillan (for Trade Policy Res. Centre), 1976.

———. "Scholars as Public Adversaries: The Case of Economics." In [A.23]. Also in *Controversies and Decisions: The Social Sciences and Public Policy,* edited by Charles Frank. New York: Sage, 1976.

———. "Some Comments on a National Economic Policy." In *Policy Formation in an Open Economy,* vol. 1, edited by Robert A. Mundell and B. E. van Snellenberg. Waterloo: Univ. Waterloo.

———. "Some Comments on Inflation Theory." In *Policy Formation in an Open Economy,* vol. 1, edited by Robert A. Mundell and B. E. van Snellenberg. Waterloo: Univ. Waterloo.

———. "Summing Up the Conference on the Monetary Mechanism in Open Economies." Presented in Helsinki.

———. "Unemployment and Inflation." Presented at a Social Science Research Council Conference.

———. "What Passes for Economics in the English Establishment." *Banker* 125 (October): 1159–61. Also in [A.27].

———. "What We Need to Know: Panel Discussion." In *International Trade and Finance: Frontiers for Research,* edited by Peter B. Kenen. Cambridge: Cambridge Univ. Press.

———. "World Inflation and the International Monetary System." *Three Banks Rev.,* no. 107 (September), pp. 3–22.

1976. "The Ability of the Current Floating Exchange Rate System to Cope over the Next Five Years." Report to the Department of State, September.

———. "Adam Smith: Lecture." Presented at the Ibero-American Univ., Mexico City, February.

———. "The American Tradition in Economics." Paper presented at the American Econ. Assoc. meetings, September. *Nebraska J. Econ. and Bus.* 16 (Summer 1977): 17–26.

———. "Aspects of Patents and Licenses as Stimuli to Innovation." Bernhard Harms Lecture, Kiel, June. *Weltwirtschaftliches Archiv* 112 (3): 417–28. Also in *Bernhard-Harms-Vorlesungen,* vol. 7, edited by Herbert Giersch. Kiel: Inst. Weltwirtschaft Univ. Kiel.

———. "Comment on Mayer on Monetarism" (with summaries in German and French). *Kredit und Kapital* 9 (2): 145–53. Also in *The Structure of Monetarism,* by Thomas Mayer et al. New York: Norton, 1978.

———. Comment on Michael G. Porter's "International Financial Integration: Long-Run Policy Implications." In *Money and Finance in Economic Growth and Development: Essays in Honor of Edward S. Shaw,* edited by Ronald I. McKinnon. New York and Basel: Dekker.

———. "Comment on My Volkov's Paper." Paper presented at the Eighth Pacific Trade and Development Conference, July.

———. "U.S. Tariff Effects on Trade and Employment in Detailed SIC Industries: Comment." Department of Labor conference, December. In *The Impact of International Trade and Investment on Employment*, edited by William Dewald, Harry Gilman, Harry Grubert, and Larry Wipf. Washington: U.S. Dept. Labor, Bureau Internat. Labor Affairs, 1978.

———. "Concluding Remarks." In *Capital Movements and Their Control*, edited by Alexander K. Swoboda. Leiden: Sijthoff (for Grad. Inst. Internat. Studies).

———. "Contemporary Problems of Industrialization of Developing Countries." Lecture presented to the Indian Chamber of Commerce, Calcutta, December.

———. "Destabilizing Speculation: A General Equilibrium Approach." *J.P.E.* 84 (February): 101–8.

———. Discussion of "Increasing International Economic Interdependence: The Implications for Research." *A.E.R. Papers and Proc.* 66 (May): 162.

———. "Economic Growth and the Distribution of Economic Opportunities and Benefits among Men and Nations: An Unenthusiastic Contribution towards Predicting the Intellectual Stock Market." Essay for the Ford Foundation, January.

———. "Economics and Long-Term Forecasting." In *Progress and Problems in Social Forecasting*, edited by Christopher Freeman, Marie Jahoda, and I. Miles. London: Soc. Sci. Res. Council.

———. "Elasticity, Absorption, Keynesian Multiplier, Keynesian Policy, and Monetary Approaches to Devaluation Theory: A Simple Geometric Exposition." *A.E.R.* 66 (June): [illegible]

———. "Foreign Ownership and F[illegible]
Problem." *J. Energy and Developmen*[illegible]

———. "How Good Was Keynes' Cam[illegible]
82–91. Also as "Cambridge as an [illegible]
Early 1930's: A Reconstruction fr[illegible]
Cambridge and "The General Theory[illegible]
Clark Leith. London: Macmillan, 19[illegible]

———. "The Individual and the State: Some [illegible] lems." Adam Smith Bicentennary Lecture, October. In *Adam Smith and the Wealth of Nations: 1776–1976 Bicentennial Essays*, edited by Fred R. Glahe. Boulder: Colorado Assoc. Univ. Press, 1978.

———. "[Industrial Policy: Location, Technology, Multinational Firms, Competition, and Integration of Product Markets]: Report on Group Discussion" (Working Group C, International Economic Association Conference). In *Economic Integration: Worldwide, Regional, Sectoral*, edited by Fritz Machlup. London: Macmillan; New York: Halsted.

———. "International Monetary Reform: The 'Sideways' Approach." *Boletin internac., Banco de Vizcaya* (May), pp. 1–4.

———. "Introduction." In *The Politicisation of Economic Decisions,* by Alan A. Walters. London: Internat. Inst. Econ. Res.

———. "The Johnsonian Savings Function and the Rate of Inflation: A Correction." *Econ. internaz.* 29 (February/May): 87–91.

———. "Keynes and the Developing World." *Spectator* 237 (November 6): 14–15. Also in *The End of the Keynesian Era: Essays on the Disintegration of the Keynesian Political Economy,* edited by Robert Skidelsky. London: Macmillan; New York: Holmes & Meier, 1977. Also as "Keynes and Development" in [A.27].

———. "Keynes' *General Theory:* Revolution or War of Independence?" Innis lecture presented at C.E.A., June. *Canadian J. Econ.* 9 (November): 580–94. Also in [A.27].

———. "The Man Who Turned Economics into Common Sense." *The Times* (London) (March 9), p. 6.

———. "The Monetary Approach to the Balance of Payments: Essential Concepts and Historical Origins" (with Jacob A. Frenkel). In [C.24].

———. "Money in the Open Economy: A Historical and Analytical Survey." Paper presented at a Social Science Research Council and London School of Economics Conference, April. In Italian in *Prospettive settana* 2 (April/June): 25–33.

———. "The North-South Issue." Paper presented at the Toyo Keizai Shinbun Symposium, Tokyo. In Japanese in *Contemporary Econ.* 25 (Winter): 20–27. Also in *The First World and the Third World: Essays on the New International Economic Order,* edited by Karl Brunner. Rochester, N.Y.: Center for Res. Government Policy and Bus., Univ. Rochester, 1978.

———. "A Note on Factor-Income Distribution and the Factor-Owner Production Block." *Manchester School Econ. and Soc. Studies* 44 (December): 303–8.

———. "Notes on the Classical Transfer Problem." *Manchester School Econ. and Soc. Studies* 44 (September): 211–19.

———. "Panama as a Regional Financial Center: A Preliminary Analysis of Development Contribution." *Econ. Development and Cultural Change* 24 (January): 261–86.

———. "The Pitfalls of International Monetary Reform." *Commerce* 72 (January): 26–32. As "Towards a World Central Bank." *Banker* 126 (February): 137–39.

———. "Protection and Welfare in Modern Perspective." *J. World Trade Law* 10 (March/April): 188–95.

———. "The Relevance of the *Wealth of Nations* to Contemporary Economic Policy." *Scottish J. Polit. Econ.* 23 (June): 171–76.

———. "Reserve Requirements and Monetary Control." Discussion paper no. 66, Economic Council of Canada, October.

———. "The Role of Networks of Economists in International Monetary Reform." Paper presented at a University of Pennsylvania conference, November.

———. "Social Science: The Public Disenchantment: A Symposium." *American Scholar* 45 (Summer): 340–41.

———. "A Summing Up." Paper presented to the Seventh Pacific Trade and Development Conference. In *Co-operation and Development in the Asia-Pacific Region: Relations between Large and Small Countries*, edited by Leslie V. Castle and Sir Frank Holmes. Tokyo: Japan Econ. Res. Center.

———. "Technology, Technical Progress and the International Allocation of Economic Activity." Paper presented at a Nobel Symposium, Stockholm, June. In *The International Allocation of Economic Activity: Proceedings of a Nobel Symposium held at Stockholm*, edited by Bertil Ohlin, Per-Ove Hesselborn, and Per Magnus Wijkman. London: Macmillan; New York: Holmes & Meier, 1977.

———. "The Theory of Tax and Expenditure Incidence: A Diagrammatic Analysis" (with Melvyn B. Krauss). *Public Finance* 31 (3): 340–62.

———. "Three Lectures on Inflation." Presented at the Central Bank of Uruguay, September.

———. "Trade, Development, and Dependence." *Queen's Q.* 83 (Autumn): 427–42.

———. "Trade, Growth, and Economic

sented to the Eighth Pacific Trade

July.

———. "Trade Negotiations and the

System." *Commercial Policies Issues*

Leiden: Sijthoff.

———. "A Tribute to Homer Jones."

437–38.

———. "What Is Right with Monetarism?"

(April), pp. 13–17.

———. "When Will the World Come out of Recession?" *Shell World*, no. 3 (April/May), pp. 12–13.

———. "World Inflation and the Developing Countries." *Mercado de capitais* 2 (May/August): 339–56.

———. "World Inflation, the Developing Countries, and 'An Integrated Programme for Commodities.'" *Banca Naz. Lavoro Q.*, no. 119 (December), pp. 309–35. Also in *Información comercial española*, no. 521 (January 1977), pp. 11–23. Also in *The First World and the Third World: Essays on the New International Economic Order*, edited by

Karl Brunner. Rochester, N.Y.: Center Res. Government Policy and Bus., Univ. Rochester, 1978.

1977. "Changing Views on Trade and Development: Some Reflections." In *Essays on Economic Development and Cultural Change in Honor of Bert F. Hoselitz,* edited by Manning Nash. *Econ. Development and Cultural Change* 25 (suppl.): S363–S375.

———. A Comment on D. A. Currie's "Some Criticisms of the Monetary Analysis of Balance of Payments Correction" (with A. Robert Nobay). *Econ. J.* 87 (December): 769–70.

———. "Commodities: Less Developed Countries' Demands and Developed Countries' Response." In *The New International Economic Order: The North-South Debate,* edited by Jagdish N. Bhagwati. Cambridge, Mass.: M.I.T. Press.

———. "Economics and the Radical Challenge: The Hard Social Science and the Soft Social Reality." In *Culture and Its Creators: Essays in Honor of Edward Shils,* edited by Joseph Ben-David and Terry N. Clark. Chicago: Univ. Chicago Press.

———. "The Elementary Economic Geometry of Buffer Stock Price Stabilization." *Malayan Econ. Rev.* 22 (April): 1–9.

———. "Equity, Efficiency, and Intervention." In *Essays in Honor of Dimitrios J. Delivanis,* edited by N. J. Pantazopoulos. Thessaloniki: Aristotelian Univ. Press.

———. "Monetarism: A Historic-Theoretic Perspective" (with A. Robert Nobay). *J. Econ. Literature* 15 (June): 470–85.

———. "The Monetary Approach to Balance of Payments Theory and Policy: Explanation and Policy Implications." *Economica* 44 (August): 217–29.

———. "The Monetary Approach to the Balance of Payments: A Nontechnical Guide." *J. Internat. Econ.* 7 (August): 251–68.

———. "The New International Economic Order." Selected Papers no. 49, Grad. School Bus., Univ. Chicago, January. Also in *Boletin internac.*, no. 5 (May), pp. 4–9. Also in *The First World and the Third World: Essays on the New International Economic Order,* edited by Karl Brunner. Rochester, N.Y.: Center Res. Government Policy and Bus., Univ. Rochester, 1978.

———. "A Note on the Dishonest Government and the Inflation Tax." *J. Monetary Econ.* 3 (July): 375–77.

———. "A Note on the Geometry of the Two-by-Two General Equilibrium Model." *Economica* 44 (February): 71–75.

———. "A Note on the Theory of Public Goods and 'Bluffing.' " *J. Public Econ.* 7 (June): 383–86.

———. "Panel Discussion on the New International Economic Order." In *The New International Economic Order: The North-South Debate,* edited by Jagdish N. Bhagwati. Cambridge, Mass.: M.I.T. Press.

———. "The Role of International Economics Teaching in the Analysis of International Relations." In *International Relations in a Changing World.* Fiftieth Anniversary Festschrift, Grad. Inst. Internat. Studies, Geneva. Leiden: Sijthoff.

———. "Rx for Economy?" or "Schools Should Train for Jobs." *Toronto Globe and Mail* (February 19).

———. "The Shadow of Keynes." *Minerva* 15 (Summer): 201–13. Also in [A.27].

———. "The Welfare Economics of Monetary Transfers and Reversed Transfers." *J. Money, Credit and Banking* 9 (May): 259–75.

———. "World Inflation, International Monetary Reform and the Less Developed Countries." V. K. Ramaswami Memorial Lecture, Delhi School of Economics, January. *Indian Econ. Rev.* 21 (April): 1–14.

1978. "An Analysis of the Economic Determinants of Legal and Illegal Mexican Migration to the United States" (with Arturo C. Porzecanski and Mario I. Blejer). In *Research in Population Economics,* vol. 1, edited by Julian L. Simon. Greenwich, Conn.: JAI

———. "Disguised Unemployment in a General Theoretical Context." *Econ. Development and Cultural Change* 26 (January): 385–89.

———. "James Meade's Contribution to Economics." *Scandinavian J Econ.* 80 (1): 64–85

———. "Ruth Cohen: A Neglected Contributor to Contemporary Capital Theory." In [A.27].

———. "Seigniorage and the 'Inflation Tax' [illegible] ting." In *Redistribution through the* [illegible] *nomics of Money and Credit,* edit[illegible] Thomas F. Wilson. New York: Pr[illegible]

1979. "Monetary and Economic Prob[illegible] The Caribbean Case." *Zeitschrift fur* [illegible] *ten,* nos. 1–2, pp. 237–41.

1981. "Factor Intensities and the Shape [illegible] Curve" (with José S. Lizondo and Yeong-[illegible] (May): 199–202.

E. Journalism

1959. "The Kind of Budget I Would Like to See." *Financial Times* (March 23).

1965. "Trade and Aid for Poorer Nations." *The Times* (December 8).

1967. "After the Kennedy Round." *Listener* (August 24).

———. "Monetary Reform." *Listener* (November 23).

1968. "Argument over Management of Money Supply" (with Brian Griffiths). *The Times* (October 29).

———. "Dethroning Gold." *Listener* (March 21).

———. "An Economist Looks at Student Unrest." *Sunday Telegraph* (December 1).

———. "Exploitation?" *Listener* (February 8).

———. "The Safety Valve in a Floating Exchange Rate." *The Times* (December 9).

1969. "Foreign Equity Investment: The Canadian Experience." *Sydney Morning Herald* (October 14).

———. "Payoff from 'Curiosity' Research." *Science J.* 5A (December): 3.

1970. "Inflation: The Textbook Gives No Answer" (with David Metcalf and Ray Richardson). *Financial Times* (November 14).

———. "Opening Up the Bank" (with David R. Croome). *The Times* (April 9).

———. "Too Many Actors" (two letters on actors' wages). *The Times* (April 8 and 16).

———. "Wage Rise Curbs in Context." *The Times* (November 23).

1971. Letter on Trade Policy Initiative. *The Times* (June 29).

———. "No Need for This Economic Gloom" (with Frank Paish). *Sunday Times* (March 7).

———. ["The Principles of a Reformed International Monetary System"] (in German). *Wirtschaftswoche der Volkswirt Aktionan* 25 (October 1).

1972. "The Budget, 1972." *Spectator* (March 25).

1973. "Cheerfulness Keeps Breaking In." Morris Wigram *Newsletter* ("Is Keynes Dead?" issue) (August), pp. 5–7.

———. "Experts Discuss Dollar Devaluation." *Yale Daily News* (February 15).

———. "International Capital Movement Controls." *Euromoney* (August), pp. 27, 29.

———. "Learning to Love Floating Rates." *New York Times* (March 18).

———. Letter on "Rich and Poor Nations." *The Times* (March 18).

———. "The Problems of Central Bankers in a World of Floating Rates." *Euromoney* (July), pp. 4–7. Also in *UMBC Econ. Rev.* 9 (2): 67–74. Also in *Econ. Notes: Monte dei Pasche di Siena* 2 (September/December): 7–19.

———. "Time to Reconsider Support for Graduate Work." *Times Higher Educ. Suppl.* (July 13).

1974. ["Economic Development Policies: The Contrasting Recent Experiences of Chile and Brazil"] (with Larry A. Sjaastad; in German). *Wirtschaftswoche* 28 (February 1): 54–58.

———. "Expert View no. 10 in Part III." In *Inflation: How to Curb Public Enemy Number One*, edited by A. Rupert. London: Inst. Econ. Affairs.

———. "Future Bright for U.S.A.: World Slump Won't Come." *Oriental Econ.*, nos. 1 and 2 (February), pp. 28–29.

———. "L'Inflation: Analyse et politique." *Bull. Inst. Econ. Paris* (December), pp. 2–3.

———. "The Western Model of Economic Development." *World Development* 2 (February): 9–10. Also in *Alternatives in Development*, edited by J. West. New York: Pergamon.

1975. "The ABC of SDRs." *Banker* 125 (December): 1489–91. Also in French in *Bull. Inst. Econ. Paris* (October), pp. 1–4.

———. "The Canadian Dilemma." *Euromoney* (December), p. 77.

———. "Central Bankers' Self-Interest Erodes Floating Exchange Rates." *Chicago Tribune* (September 8).

———. ["Have Floating Rates Worked So Badly?"] (in French). *Bull. Inst. Econ. Paris* (July/August), pp. 1–5.

———. "IMF Hit as Failing to Provide World Money Leadership." *Chicago Tribune* (March 24).

———. "Let Arab Play Business Game Here by Our Rules." *Chicago Tribune* (June 16).

———. "Little Joy in Jolly Old England." *Chicago Tribune* (January 13).

———. "1975 or 1984?" In *Crisis '75 . . .?* Occasional Paper Special no. 43. London: Inst. Econ. Affairs.

———. "UN-Style Freedom a Bad Joke." *Chicago Tribune* (October 20).

1976. "Friedman's Nobel Laureate." *London Economist* (October 23)

———. "New Reasons for Optimism about Growth " *Davy Rev*

———. "Poor Likely to Lose in Pursuit of 'New World Trade Order.' " *Chicago Tribune* (January 19).

F. Miscellaneous

1951. "The Economics of Undertak [illegible] 240–44.

1954. "The Private Eye of Mickey S [illegible] 16–18.

1965. "The Progressive Society." Convocation Address, St. Francis Xavier Univ., May.

1967. "The Problems of the World Economy." Convocation Address, Queen's Univ., May.

1969. "Autobiographical Notes." Lakehead Univ., June.

———. "Myth and Reality." *Univ. Sheffield Gazette* 19 (November) 17–19.

1970. "Canada and Contemporary Society " Convocation Address, Carleton Univ., May.

———. "The Economic Future of Sex." In *Visions 2020*, edited by G. Clarkson. Toronto: Canadian Forum.

———. "Guidelines for Relationships between Governments and Multinational Corporations." Paper presented at a meeting of the British–North American Committee, Palm Beach, Fla.

———. "Report of the Committee on the Criteria of Academic Appointment." Univ. Chicago *Record* (December 17), pp. 1–25.

1971. Economists' letters pro and con joining the Common Market. *The Times* (October 22).

———. "From Gordon to Watkins to Uselessness." Address to Ontario Institute of Management Consultants, May.

———. Letter on the award of the 1971 Nobel Prize in Economics to Simon Kuznets. *The Times* (October 26).

1973. "Monetary Ordeal." *The Times* (October 18).

1974. "Cambridge in the 1950s: Memoirs of an Economist." *Encounter* 42 (January): 28–39. Also in [A.23; A.27].

———. "Me and My Brain Drain." *Punch* (October 30).

1975. "Inflation, Energy, and the World Economy: Prospects for International Cooperation." Panel statement presented at the Quadrangular Conference II, Washington.

———. "J.A.P.E.—Jokes with a Point for Economists." *Eastern Econ. J.*, vol. 2 (January, April, June).

———. Speech presenting sculpture to the University of Western Ontario. October 10.

1978. "Are Savings Male or Female?" In [A.27].

G. Book Reviews, Signed and Unsigned

1949. Jellinek, Frank. *West of the Decline*. *Econ. J.* 59 (June): 289.

———. Maxwell, James A. *The Fiscal Impact of Federalism in the United States*. *Econ. J.* 59 (September): 432–34.

———. Wiener, Norbert. *Cybernetics*. *Econ. J.* 59 (December): 573–75.

———. Williams, John H. *Postwar Monetary Plans*. *Econ J.* 59 (September): 487.

1950. Anderson, Benjamin M. *Economics and the Public Welfare: Financial and Economic History of the United States, 1914–1946*. *Econ. J.* 60 (December): 819–21.

———. Anderson, Montgomery D. *Total Business Fluctuations*. *Econ J.* 60 (June): 442.

———. Busschau, William J. *The Measure of Gold: The Role of Gold as International Money*. *Econ. J.* 60 (September): 571–73.

———. Daugherty, Carroll R., and Daugherty, Marion R. *Principles of Political Economy*. *Economist* 159 (July 8): 72.

———. Duesenberry, James S. *Income, Saving and the Theory of Consumer Behaviour. Economist* 159 (July 15): 119.

———. Hobson, John A. *The Science of Wealth. Econ. J.* 60 (September): 569–71.

———. James, R. Warren. *Wartime Economic Co-operation: A Study of the Relation between Canada and the United States. Canadian J. Econ. and Polit. Sci.* 16 (August): 435–38.

———. Kimmel, Lewis H. *Taxes and Economic Incentives. Economist* 159 (September 16): 476.

———. Moulton, Harold G. *Controlling Factors in Economic Development. Econ. J.* 60 (March): 135–37.

———. Smith, R. Elberton. *Customs Valuation in the United States. Econ. Hist. Rev.* 2 (3): 342.

———. Taylor, Horace, and Barger, Harold. *The American Economy in Operation;* and Harris, Seymour E., ed. *Foreign Economic Policy for the United States. Econ. J.* 60 (June): 374–77.

———. Wilson, Thomas. *Modern Capitalism and Economic Progress;* and Keezer, Dexter M., et al. *Making Capitalism Work. Economist* 159 (July 1): 16.

1951. Beyen, J. W. *Money in a Maelstrom. Econ. J.* 61 (September): 600–601.

———. Boulding, Kenneth E. *A Reconstruction of Economics Economist* 160 (February 17): 369.

———. Brems, Hans. *Product Equilibrium under Monopolistic Competition. Economist* 161 (December 8): 1399

———. Bye, Raymond T. *Social Econo*

in Welfare Economics. Economist 160

———. Chang, Tse Chun. *Cyclical M*

Economist 161 (August 4). 283.

———. Clemence, Richard V., and I

ian System. Econ. J. 61 (March): 13

———. Dunlop, John T. *Wage Deter*

omist 160 (February 17): 370.

———. Korner, Emil. *The Law of Freedom as the Remedy for War and Poverty. Economist* 161 (September 22): 678.

———. Lutz, Friedrich, and Lutz, Vera. *The Theory of Investment of the Firm. Economist* 161 (December 29): 1577.

———. Marsh, Donald B. *World Trade and Investment. The Economics of Interdependence. Economist* 160 (July 14): 85.

———. Mints, Lloyd W. *Monetary Policy for a Competitive Society Econ J.* 61 (June): 382–84.

———. Mitchell, Wesley C. *What Happens in Business Cycles A Progress Report. Economist* 161 (August 18): 394–95

———. Thorne, William J. *Banking. Econ. J.* 61 (March): 227.

———. Tinbergen, Jan, and Polak, J. J. *The Dynamics of Business Cycles: A Study in Economic Fluctuations. J. Royal Statis. Soc.* 114:260–61.

1952. American Economic Assoc. *Readings in Monetary Theory. Economist* 165 (November 8): 377.

———. Chipman, John S. *The Theory of Inter-sectoral Money Flows and Income Formation. Economica* 19 (November): 436.

———. Condliffe, John B. *The Commerce of Nations. Cambridge J.* 5 (April): 448–50.

———. Dacey, W. Manning. *The British Banking Mechanism. Econ. J.* 62 (June): 381–83.

———. Feis, Herbert T. *The Diplomacy of the Dollar. Economist* 162 (March 29): 790.

———. Gruber, Karl. *Conditions of Full Employment. Economist* 163 (May 10): 369.

———. Klein, Lawrence R. *Economic Fluctuations in the United States, 1921–41. Internat. Affairs* 28 (April): 216–17.

———. Sen, S. N. *Central Banking in Undeveloped Money Markets. Econ. J.* 62 (December): 893–96.

———. Stuvel, Gerhard. *The Exchange Stability Problem. Economist* 162 (January 12): 82.

1953. Higgins, Benjamin H. *What Do Economists Know? Six Lectures on Economics in the Crises of Democracy. Canadian J. Econ. and Polit. Sci.* 19 (May): 270–72.

———. Sraffa, Piero, and Dobbs, M. H., eds. *Works and Correspondence of David Ricardo. Cambridge J.* 6 (April): 444–46.

———. Wright, D. M., ed. *Money, Trade and Economic Growth: In Honor of John Henry Williams. Econ. J.* 63 (December): 855–57.

1954. Burns, Arthur F. *The Frontiers of Economic Knowledge. A.E.R.* 44 (December): 939–41.

———. Neisser, Hans, and Modigliani, Franco. *National Incomes and International Trade. Economica* 21 (August): 264–65.

———. Sayers, Richard S. *Banking in the British Commonwealth. Econ. J.* 64 (March): 149–52.

1955. Dehem, Roger. *L'Efficacité sociale du système économique. Canadian J. Econ. and Polit. Sci.* 21 (February): 102–4.

———. Fulford, Roger. *Glyn's, 1753–1953: Six Generations in Lombard Street. Canadian J. Econ. and Polit. Sci.* 21 (May): 256–58.

———. Haley, B. F., ed. *A Survey of Contemporary Economics, Vol. II. Econ. J.* 65 (June): 302–6.

———. Marcus, Edward. *Canada and the International Business Cycle, 1929–1939. Econ. J.* 65 (December): 749.

1956. Brown, Arthur J. *The Great Inflation, 1939–1951. Econ. J.* 66 (March): 121–23.

———. Ferns, H. S., and Ostrey, B. *The Age of Mackenzie King: The Rise of the Leader. Econ. J.* 66 (June): 398.

———. Haberler, Gottfried. *A Survey of International Trade Theory. Econ. J.* 66 (June): 402–3.

———. Hawtrey, Ralph G. *Towards the Rescue of Sterling. Econ. J.* 66 (June): 336–37.

———. Hicks, John R. *A Revision of Demand Theory. Manchester Guardian* (August 3).

———. International Economic Assoc. *International Economic Papers, no. 5. Econ. J.* 66 (September): 506–7.

———. Meade, James E. *A Geometry of International Trade. Econ. J.* 66 (June): 319–21.

———. Neufeld, Edward P. *Bank of Canada Operations, 1935–54. Canadian Forum* 36 (July): 91–92.

———. Robertson, Dennis H. *Economic Commentaries. Manchester Guardian* (August 3).

1957. Allen, R. G. D. *Mathematical Economics Manchester Guardian* (June 21).

———. Bell, Philip W. *The Sterling Area in the Postwar World Econ Hist. Rev.*, vol. 10 (August).

———. de Van Graaff, J. *Theoretical Welfare Economics Manchester Guardian* (June 28).

———. Fellner, William. *Trends and Cycles in Economic Activity A.E.R.* 47 (March): 164–66.

———. Scott, A. D. *Natural Resources: The Economics of Conservation. Economica* 24 (February): 80–81.

1958. Galbraith, John Kenneth. *The Affluent Society. Spectator* 201 (September 19): 381.

———. Gillman, Joseph M. *The Falling Rate of Profit Accounting Res* 9 (October): 340–41.

———. Harrod, Roy. *Policy against In* [illegible] 24): 556–57.

———. Kelf-Cohen, Reuben. *Nationa* [illegible] *Dogma. Spectator* 201 (December 12) [illegible]

———. Makower, Helen. *Activity And* [illegible] *Equilibrium. Economica* 25 (May): 155 [illegible]

———. Roberts, B. C. *National Wages Polic* [illegible] Davison, J. P.; Florence, P. Sargent; and Gram, Barbara [illegible] *ity and Economic Incentives. Spectator* 201 (July 25) 144

———. Thomas, Brinley. *The Economics of International Migration*, and Strauss, Erich. *Common Sense about the Common Market. Spectator* 200 (May 2): 568.

———. Wicksell, Knut. *Selected Papers on Economic Theory Manchester Guardian* (April 11).

1959. Jackson, Barbara Ward. *Five Ideas That Changed the World. Spectator* 202 (March 20): 414.

———. Maxey, George, and Silberston, Aubrey *The Motor Industry;*

Cook, P. Lesley. *Effects of Mergers;* Hagenbuch, Walter. *Social Economics;* and Marris, Robin. *Economic Arithmetic. Spectator* 202 (February 13): 235.

———. Mills, C. Wright. *The Sociological Imagination. Spectator* 203 (September 4): 310.

———. Murray, David. *Steel Curtain. Spectator* 203 (July 10): 45–46.

———. Scitovsky, Tibor. *Economic Theory and Western European Integration. A.E.R.* 49 (December): 1103–5.

1960. Hicks, John R. *Essays in World Economics. Economica* 26 (August): 279–80.

———. Kierstead, B. S. *Capital, Interest, and Profits. J.P.E.* 68 (October): 525–26.

———. Mookerjee, Subimal. *Factor Endowments and International Trade: A Study and Appraisal of the Heckscher-Ohlin Theory. J.P.E.* 68 (August): 430–31.

———. Sannwald, Rolf F., and Stohler, Jacques. *Economic Integration: Theoretical Assumptions and Consequences of European Integration. J.P.E.* 68 (April): 206–7.

———. Vaccara, Beatrice N. *Employment and Output in Protected Manufacturing Industries. Current Econ. Comment* 22 (November): 68–69.

1961. Kilbourn, William. *The Elements Combined: A History of the Steel Company of Canada. Internat. J.* 16 (Autumn): 425–26.

———. Letiche, John M. *Balance of Payments and Economic Growth. Canadian J. Econ. and Polit. Sci.* 27 (February): 134.

———. Meade, James E. *A Neo-Classical Theory of Economic Growth. J.P.E.* 69 (October): 498–500.

1962. Aitken, Hugh G. J. *American Capital and Canadian Resources. J.P.E.* 70 (August): 118–19.

———. Brewis, T. N.; English, H. E.; Scott, Anthony; and Jewett, Pauline. *Canadian Economic Policy. Canadian J. Econ. and Polit. Sci.* 28 (February): 161–62.

———. Frost, Raymond. *The Backward Economy. Queen's Q.* 69 (Autumn). 476.

———. Lutz, Friedrich. *The Problem of International Economic Equilibrium. Canadian J. Econ. and Polit. Sci.* 28 (November): 639.

———. Sraffa, Piero. *Production of Commodities by Means of Commodities: Prelude to a Critique of Economic Theory. Canadian J. Econ. and Polit. Sci.* 28 (August): 466.

1964. Geertz, Clifford. *Peddlers and Princes. J.P.E* 72 (February): 104–5.

———. Janssen, L. H. *Free Trade, Protection and Customs Union. J.P.E.* 72 (April): 208–9.

———. Linder, Staffan B. *An Essay on Trade and Transformation. Economica* 31 (February): 86–90.

———. Machlup, Fritz. *Essays on Economic Semantics. Canadian J. Econ. and Polit. Sci.* 30 (May): 294–96.

———. Vernon, Raymond. *The Dilemma of Mexico's Development. Queen's Q.* 71 (Summer): 280–81.

1965. Curzon, Gerard. *Multilateral Commercial Diplomacy: The General Agreement on Tariffs and Trade and Its Impact on National Commercial Policies and Techniques. J.P.E.* 73 (December): 659–60.

———. Economic Council of Canada. *First Annual Review Economic Goals for Canada to 1970. J.P.E.* 73 (August): 410–12.

———. Kemp, Murray C. *The Pure Theory of International Trade Canadian J. Econ. and Polit. Sci.* 31 (February): 139–41.

1966. Balassa, Bela A. *Trade Prospects for Developing Countries. Econ Development and Cultural Change* 14 (January): 247–49

———. Pincus, John. *Economic Aid and International Cost Sharing. J.P.E.* 74 (February): 94–95.

1967. Boyle, Andrew. *Montagu Norman: A Biography. Listener* 78 (November 9): 609.

———. Galbraith, John Kenneth. *The New Industrial State. Spectator* 219 (November 10): 574–76.

———. Lekachman, Robert. *The Age of Keynes. Spectator* 218 (June 23): 740–41.

———. Patterson, Gardner. *Discrimination in International Trade: The Policy Issues 1945–65. A.E R.* 57 (September): 958–60.

———. Richardson, Harry W. *Economic Recovery in Britain, 1932–39 Director* 19 (June): 500–503.

———. Robinson, Joan. *Economics: An*
Graham. *Workers, Unions and the Stat*
15–16.

1968. Cooper, Richard N. *The Econo*
Policy in the Atlantic Community. Spect

———. Denison, Edward F. *Why Grow*
122–24.

———. Janeway, Eliot. *The Economics of*
Dollar. Spectator 221 (July 19): 93–94.

———. Leontief, Wassily W. *Essays in Economics. Econ J* 78 (September): 664–66.

———. Servan-Schreiber, Jean-Jacques. *Le Defi americain Interplay* (June/July), pp. 56–59.

———. Triffin, Robert. *Our International Monetary System. Book World* 2 (August 11): 19.

1969. Caves, Richard E., et al *Britain's Economic Prospects. Polit. Q.* 40 (January/March): 121–25.

———. Dunning, John H. *Role of American Investment in the British Economy. Director* 21 (May): 318–21.

———. Lutz, Vera. *Central Planning for the Market Economy. Director* 20 (April): 132.

1970. Bhagwati, Jagdish N. *Trade, Tariffs and Growth: Essays in International Economics. J.P.E.* 78 (November/December): 1380–81.

———. Claassen, Emil M. *Monnaie, revenu national et prix. J.P.E.* 78 (July/August): 793–94.

———. Kindleberger, Charles P. *Power and Money: The Politics of International Economics and the Economics of International Politics. Spectator* 225 (October 31): 524–25.

———. Scitovsky, Tibor. *Money and the Balance of Payments. J.P.E.* 78 (March/April): 424–25.

———. Thomas, D., ed. *The Mind of Economic Man. Director* 23 (December): 560.

1971. Baldwin, Robert E. *Nontariff Distortions of International Trade. Monthly Labor Rev.* 74 (June): 94–95.

———. Buchanan, James M., and Devletoglou, Nicos E. *Academia in Anarchy. J.P.E.* 79 (January/February): 203–5.

———. Devons, Ely. *Papers on Planning and Economic Management.* Edited by Alec Cairncross. *Economica* 38 (May): 204–5.

———. Lewis, W. Arthur. *Aspects of Tropical Trade, 1883–1965. J. Internat. Econ.* 1 (February): 131–36.

———. Stykolt, Stefan. *Efficiency in the Open Economy: Collected Writings on Canadian Economic Problems and Policies.* Edited by Anthony Scott and James D. Rae. *Economica* 38 (May): 220–21.

1972. Lipsey, Richard G. *The Theory of Customs Unions: A General Equilibrium Analysis. Econ. J.* 82 (June): 728–30.

———. Miller, Roger L., and Williams, Rayburn M. *The New Economics of Richard Nixon;* and Silk, Leonard. *Nixonomics. New York Times Book Rev.* (August 27).

———. Robbins, Lloyd. *Autobiography of an Economist. J.P.E.* 80 (July/August): 835–36.

———. Robbins, Lloyd. *Money, Trade and International Relations. J. Internat. Econ.* 2 (August): 307–9.

1973. Kalecki, M. *Selected Essays in the Dynamics of the Capitalist Economy. Economica* 40 (February): 87–88.

1974. Batra, Raveendra N. *Studies in the Pure Theory of International Trade. Economica* 41 (November): 462–63.

———. Boulding, Kenneth E. *The Economy of Love and Fear: A Preface to Grants Economics. J.P.E.* 82 (September/October): 1055–56.

———. Brittan, Samuel. *Is There an Economic Consensus? An Attitude Survey. Economica* 41 (May): 214–16.

———. Despres, Emile. *International Economic Reform: Collected Papers.* Edited by Gerald M. Meier. *J. Internat. Econ.* 4 (August): 317–18.

———. Friedman, Milton. *Money and Economic Development: The Horowitz Lecture of 1972. Economica* 41 (August): 346–47.

———. Lieftinck, Pieter. *The Post-War Financial Rehabilitation of the Netherlands. Finance and Development* 11 (June): 33.

———. Robinson, Ronald, ed. *Developing the Third World: The Experience of the Nineteen-Sixties. Polit. Sci. Q.* 89 (Fall): 709–11.

1975. Blaug, Mark. *The Cambridge Revolution: Success or Failure? J.P.E.* 83 (October): 1083–84.

———. Fogel, Robert W., and Engerman, Stanley L. *Time on the Cross: The Economics of American Negro Slavery. Encounter* 44 (January): 56–59.

———. Hicks, John R. *The Crisis in Keynesian Economics. J.P.E.* 83 (June): 671–73.

———. Hicks, John R. *The Crisis in Keynesian Economics;* Moggridge, D. E., ed. *Keynes: Aspects of the Man and His Work;* and Tobin, James. *The New Economics One Decade Older. Times Literary Suppl.* (February 7).

———. Mason, Edward S., and Asher, Robert E. *The World Bank since Bretton Woods. Economica* 42 (November): 445–46.

1976. Goodheart, C. A. E. *Money, Information and Uncertainty. Econ. J.* 86 (December): 884–86.

———. Mandel, Ernest. *Late Capitalism;* and Carré, Jean-Jacques; Dubois, Paul; and Malinvaud, Edmond. *French Economic Growth. J. Modern Hist.* 48 (December): 696–98.

———. Prachowny, Martin F. J. *Small Open Economies Their Structure and Policy Environment. J.P.E.* 84 (April): 420.

———. Salin, Pascal. *Economie internationale.* vol 1, *Economie financiere. J. Internat. Econ.* 6 (August): 309–14

1977. Bacon, Robert, and Eltis, Walter *Few Producers. Canadian Public Polic*

———. Brittan, Samuel. *Capitalism an Polit. Sci. Rev.* 71 (December): 161

———. Brown, C. P. *Primary Commodi* 4 (March): 80–83.

1978. Sen, A. K. *Employment, Technology a opment and Cultural Change* 26 (January): 38

Schooling, Search, and Spouse Selection: Testing Economic Theories of Marriage and Household Behavior

Bryan L. Boulier
George Washington University

Mark R. Rosenzweig
University of Minnesota

The implications of the economic theories of marriage and of household fertility behavior are tested in a framework in which educational investment, marital search, and marital matches are responsive to marriage-market conditions and the personal traits of individual agents, some of which are unobserved by the econometrician. Implications are also derived and tested for the effects of longevity, attractiveness, preferences, and labor- and marriage-market conditions on schooling, marriage age, and spouse choice. The empirical results indicate that inattention to heterogeneity and marital selection leads to a false rejection of the economic theory of marriage but to a false acceptance of the value-of-time fertility hypothesis.

I. Introduction

How individuals sort themselves through marriage into household units has important implications for the distribution of incomes, consumption, and fertility. Only relatively recently, however, have econo-

---

We are grateful to the referees and to Anthony Yezer for useful comments, to members of the workshops at the Economic Research Center, NORC; Miami University, University of Pennsylvania, and Yale University, and to Mercedes Concepcion for use of the 1973 National Demographic Survey.

[*Journal of Political Economy*, 1984, vol. 92, no. 4]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9204-0009$01.50

mists turned their attention to family formation, making use of (household) production and search theory (Becker 1973, 1974; Keeley 1977). One potentially significant shortcoming of this theoretical literature is the assumption, carried forward in empirical applications (Frieden 1974; Keeley 1977; Becker, Landes, and Michael 1977; Anderson and Hill 1980), that all of the traits individuals bring to the marriage market are exogenous. More precisely, empirical applications of this framework have assumed that any traits that might determine marital outcomes but that are unobserved by the researcher are uncorrelated with observed characteristics such as schooling attainment.[1] Similarly, empirical applications of demand theory have assumed that the joint characteristics of households (e.g., husband's and wife's schooling) are uncorrelated with the unobserved traits "matching" the marital partners (e.g., beauty, intelligence), which yield utility to the household and which may thus be correlated with observed consumption patterns.[2]

In this paper we attempt to test the central implications of the economic theory of marriage and of the economic theory of one consumption good, fertility, in a framework in which educational investment, marital search, and marital matches are responsive to marriage-market conditions and the personal traits of individual agents, some of which are unobserved by the econometrician. In particular, we test the propositions that there is positive assortative mating of complementary traits, that search in the marriage market yields a "better" spouse, and that women with higher levels of schooling bear fewer children because their time has a higher value. In formulating these tests we are also able to derive and test additional implications for the effects of longevity,
and labor-market and marriage-mar
tainment, age at marriage, spouse ch

The tests utilize micro data from
area-level information on life expect
and population sex ratios. The Phil
testing a model of marriage because
divorce and the low percentage of married

---

[1] Sociological research on the marriage market, based on the exchange of spouse attributes (e.g., Elder 1969; Taylor and Glenn 1976), has examined how the parental background and schooling attainment of women as well as physical attractiveness are correlated with the socioeconomic status of their husbands. Such studies have tended, however, to neglect marital search behavior and characteristics of the marriage market and also to treat all traits as exogenous

[2] Nerlove (1974) has suggested that the interpretation of household demand relations is made difficult when marital and household decisions are not recursive. His hypothesis that heterogeneity in preferences for children is responsible for the oft-observed negative relationship between female education and fertility is tested below

the home (less than 5 percent). The strength and pervasiveness of the division of labor between spouses thus make it possible to test for positive assortative mating of complementary traits based on correlations of observed schooling levels between spouses. Moreover, despite the low levels of labor-force participation by married women (Rosenzweig 1976), schooling levels of women in the Philippines are almost equal to those of men, a fact that cannot be readily explained by conventional human capital models that ignore the marriage market.[3]

The empirical results indicate that inattention to heterogeneity in unobserved marriage-market traits and to the responsiveness of schooling to marriage-market conditions leads to a false rejection of the economic theory of marriage but to a false acceptance of the value-of-time fertility hypothesis based on education-fertility correlations in the Philippines. The results reject the hypothesis that the principal cause of the latter is a negative correlation between schooling investment and "tastes" for children, but suggest, consistent with the marital selection framework, that educational levels and the age at marriage and fertility of women in the Philippines are functions of women's "marriageability," such that exogenously less attractive women receive more schooling, marry later, and have higher fertility, ceteris paribus, than do more attractive women.

II. The Mating Function: Heterogeneity and Spouse Selection

The central propositions of the theory of marriage, formulated to explain both the set of marital matches that will occur in an economy consisting of heterogeneous agents and the life-cycle timing of family formation for an individual, are (1) the positive assortative mating of agents with respect to complementary characteristics (Becker 1973) and (2) the existence of payoffs to search in the marriage market (Keeley 1977). That is, postponement of marriage results in a "better" spouse. In an economy where the household division of labor between market and nonmarket activities across mates is strong, moreover, education is presumed to be a complementary factor in marital "output." In such an environment, there will exist a structural relationship—the mating function—analogous to the individual (or own) earnings function, which relates an individual's endowments, including formal schooling, and search time, as well as the characteristics of the marriage and labor markets to the human capital, earnings potential, or other traits of the spouse obtained through search in the marriage market.

---

[3] Behrman and Wolfe (1983) discuss the potential importance of ignoring marriage-schooling interrelationships in assessing the returns to schooling.

In linear form, the mating function for individual i of sex K is

$$H^{ji} = \beta_0^K + \beta_1^K T_E^i + \beta_2^K T_S^i + \beta_3^K \alpha^i + \beta_1^K D^l + \beta_5^K R^K + \beta_6^K J^K + \beta_7^K J^l + \epsilon_1^{Ki}, \tag{1}$$

K = male, l = female; K = female, l = male; where H^{ij} is the human capital of the spouse j of individual i, T_E^i and T_S^i are i's schooling attainment and marital search time, α^i represents other traits of i that attract a higher-quality mate (marital attractiveness), J^K and J^l are variables representing sex-specific labor-market conditions, D and R are features of the marriage market reflecting the dispersion of potential mates and the ratio of potential mates to competitors (Frieden 1974), and ϵ^i is the random error term. The marital literature suggests that β_1, β_2, $\beta_5 > 0$, and $\beta_1 < 0$, while $\beta_3 > 0$ by definition.

Equation (1) may be generalized to include as a right-hand-side variable any transfers, conditional on the marriage of i and j, provided to j by the family of individual i (dowry, if i is female; bride price, if i is male). In addition, if individual i's parental background, for given observed characteristics of i, affects marital offers, then variables describing i's family would be included in the mating function.[1]

Whether least squares yields unbiased estimates of the parameters of the mating function depends on whether schooling and search time are uncorrelated with the error te [illegible]
If, for example, any traits of i that in [illegible]
attracted (e.g., beauty, intelligence) ar [illegible]
estimates of the β's will be biased an [illegible]
possibly incorrect as long as such obs [illegible]
vestments or search time are influen [illegible]

Assume, for example, that parents [illegible]
finance the schooling T_E^i of their offspring a [illegible]
mates of their own choosing, with both parents and offspring taking into account the offspring's exogenously given endowment of α and the structural relationships embodied in the mating function (1).

---

[1] We ignore here the existence of traits that are unobservable to the marriage-market participants and thus the role of family background as an indicator of such traits. Participants' unobservables play a major role in the theory and empirical analyses of marital dissolution (Becker et al. 1977) as well as in income distribution (Becker and Tomes 1979). Clearly, unforeseen random events influence marital outcomes, and uncertainty may importantly affect behavior ex ante. However, information asymmetries among optimizing agents and the researcher that arise from data constraints, which can importantly bias estimates of systematic behavioral relationships, loom large in personalized markets such as the marriage market and have not been given much attention in empirical work.

which describe the marital payoffs to schooling and search. Parents also know the decision rules of the offspring. Assume also that the lifetime welfare maximized by the offspring who live Ω years is the weighted sum of per period utilities in the postschool, unmarried, and married states, the weights being the time in each state, and that per period utilities in the postschool, unmarried, and married states are each a function of own schooling attainment. Finally, employment opportunities for *i* affect premarital income or utility, while postmarital utility is a function of the job opportunities and human capital of *i* and *i*'s spouse. Then, in linear form, the schooling attainment equation for *i* is

$$\begin{aligned} T_E^i = \gamma_0^K + \gamma_1^K\Omega' + \gamma_2^K J^K + \gamma_3^K J^l + \gamma_4^K D^l + \gamma_5^K R^K \\ + \gamma_6^K P_z + \gamma_7^K P_E + \gamma_8^K I' + \gamma_9^K \alpha' + \epsilon_2^{K\prime}, \end{aligned} \tag{2}$$

where P_z and P_E are price indices for parental consumption goods and schooling, I' is the income of *i*'s family, and J^K and J^l are variables characterizing job opportunities for *i* and for *i*'s potential mates.

Parental schooling investment decisions thus depend on the exogenous variables describing the son or daughter, the labor market, the availability of mates, schooling costs, and the family budget constraint. The search time chosen by offspring will depend on the parameters α, Ω, J, D, and R as well as on the predetermined schooling level provided by parents; that is, the (linearized) marital search equation is

$$\begin{aligned} T_S^i = \psi_0^K + \psi_1^K T_E^i + \psi_2^K\Omega' + \psi_3^K J^K + \psi_4^K J^l \\ + \psi_5^K D^l + \psi_6^K R^K + \psi_7^K \alpha' + \epsilon_3^{K\prime}. \end{aligned} \tag{3}$$

If the attractiveness endowments α are randomly distributed in the population and are unobserved by the econometrician, the estimating equation error terms in (1) through (3) will contain both the random errors ϵ and the attractiveness "effects" associated with α'. Equations (1), (2), and (3) thus form a nonrecursive system, and single-equation estimates of (1) (and of [2]) will be biased, leading to possible misleading conclusions about the theory.

None of the empirical studies of marriage in the economics literature has attempted to estimate the mating function or to take into account heterogeneity. Instead, age at marriage of the wife (search time) is treated as a dependent variable and her and her husband's characteristics are used as exogenously determined explanatory variables (Keeley 1977; Anderson and Hill 1980). These specifications thus confound the parameters of the search-time and mating equations (1) and (3). It is assumed, not only that the education of the married woman is exogenous (orthogonal to all other relevant traits), but also that her choice of a mate, in terms of his characteristics, is "exogenous," even though in both studies the objects of marital search are presumably spouse characteristics.

Consistent estimates of the relations implied by the theory between observables in the presence of heterogeneity can be obtained, however, since rank and order conditions for identification are satisfied in this case. Identification of the parameters of theoretical interest is in part achieved by the assumption that family resources, controlling for the offspring's endowed personal characteristics, have no direct effect on the quality of the mate obtained. If schooling costs (P_k) or the price of consumption goods (P_z) do not vary in the data, moreover, this assumption is necessary for identification. However, in a setting in which there is a strong household division of labor such that few women, say, are employed outside the home, and thus where positive assortative mating with respect to schooling is expected, the mating functions will not be symmetric across sex groups. Specifically, while the extent of job opportunities or employment conditions for men will affect the quality of the mates men can attract, changes in job opportunities for women will not affect marital income and thus will not directly affect the quality of the marital offers women receive from men, given women's search time. In such a setting $\beta_0^K = 0$, K = female, and the exclusion of J^i from the mating function for women is an overidentifying restriction that can be tested

III. Marital Attractiveness, Schooling

Since testing the propositions of the ec
population heterogeneous in unobse
about the determination of schooling
useful to be able to test at least some of
from those assumptions and the basic th
framework that yields equations (2) and (3), g
(1), is depicted by two subproblems. First, offspring i, endowed with a
level of attractiveness α^i and provided an optimal level of schooling
selects an optimal search time (and thus an "optimal" mate) by solving

$$\max W^i = (T_s - T_k^i)S^i + (\Omega - T_s^i)M^i \tag{4}$$

subject to the mating function (1), where S^i = postschool, "single" income, or utility per period, and M^i = per period marital income or utility accruing to offspring i. Per period single and marital welfare are described by the functions $S^{iK} = S(T_k^i, \alpha^i, J^K)$ and $M^{iK} = M(H^{ij}, T_k^i, \alpha^i, J^i, J^K)$, where, presumably, $S_1, S_2, S_3 > 0$; $M_1, M_2, M_3, M_4 > 0$ That is, own schooling and the attractiveness endowment contribute positively to both single and marital welfare, own job opportunities *contribute positively to single welfare, and spouse's human capital and job-market opportunities contribute positively to marital welfare.*

Parents of i provide an optimal schooling level to offspring by solving

$$\max_{Z, T_E^i} U = U[Z^i, W(\alpha^i)] \tag{5}$$

subject to the family budget constraint (6),

$$I^i = P_z Z^i + P_E T_E^i, \tag{6}$$

and with knowledge of offspring behavior, as characterized by (4).

The linearized solution to the offspring's problem is equation (3); the solution to the parents' problem is equation (2). The necessary first-order condition for the offspring, suppressing superscripts, is

$$(\Omega - T_s)\beta_2 M_1 = M - S = G. \tag{7}$$

Expression (7) indicates that the optimal duration of search time is determined when the gains from marriage per unit of time, G (presumed positive), are equal to the marital search gains accruing over the duration of the marriage. The effects of changes in the exogenous marriage-market and labor-market variables as well as the endowments of schooling and α will thus affect search time depending on how they are presumed to affect the gains from search and the difference between welfare in marriage and welfare when single. Similarly, the effects of changes in labor-market and marriage-market conditions and in the endowment α on schooling attainment will depend on how such variables affect the costs and returns to schooling, as well as how schooling affects search time. If we assume that (1) schooling raises market more than home productivity, so that an increase in schooling raises single more than marital income for women, given the household division of labor, as implied in Becker (1974)—that is, $dG/dT_E < 0$—while (2) attractiveness α increases the gains to being married—that is, $dG/d\alpha > 0$—and (3) time in school is incompatible with both work and marriage, while spouse search can be conducted jointly with schooling and work (so that $T_s \geq T_E$), then it can be shown that an exogenous rise in schooling will lead to a higher female marriage age, holding constant the level of the exogenous trait α, while among women with the same schooling, marriage will occur earlier in the life cycle for more attractive (higher α) women and for women residing in locations where the sex ratio is more favorable (higher R)—that is, $\psi_1 > 0$; ψ_7, $\psi_6 < 0$.[5]

---

[5] Marriage age and marital search will be monotonically related as long as the initiation date for search (puberty?) is random or the same for all agents. Termination of

Moreover, since the tendency for high α women and/or women in marriage markets with favorable demographic endowments to marry early reduces the returns to education investment (which are lower in marriage), as long as the direct effects of changes α, D, and R on the returns to schooling in marriage are small, it can be shown that less attractive women and women in unfavorable marriage markets will receive higher levels of schooling than their more attractive counterparts or than women situated in areas in which potential mates are scarcer or more dispersed; that is, $\gamma_9, \gamma_5 < 0$; $\gamma_4 > 0$. Finally, greater job opportunities for the young will lower schooling investments by increasing the opportunity cost of schooling, $\gamma_2 < 0$, and increases in life expectancy Ω will induce higher investments in both schooling and search, since the total payoffs to each will be increased, $\psi_2, \gamma_1 > 0$.

The signs of the schooling and marital search relations implied by the model also suggest the direction of the biases imparted to OLS estimates of the assortative mating and search payoff parameters in (1) due to heterogeneity in attractiveness. For example, the slope of the observed association between the ages at marriage of observationally identical (same T_k) women and the human capital of their spouses, when the women are heterogeneous with respect to α, dH/dT_s, can be written as

$$\frac{dH}{dT_s} = \beta_2 + \beta_3\left(\frac{dT_s}{d\alpha}\right)^{-1}. \qquad (8)$$

The first term in (8) is the true re
human capital of the spouse, from t
term reflects the structural effect of
tween marriage age and α. Since the
be negatively associated (more attrac
then as long as attractiveness leads to
educated potential mates, the returns to search
estimated by the association between H and T_s, conditional on all observables, if there is population heterogeneity; marriage selects out sequentially the more attractive. Of course, observed schooling levels will also be correlated with attractiveness; indeed, the partial association between the schooling levels of women and their marriage age will be an upward-biased estimate of the "true" effect of schooling on

---

search (marriage) is thus the control variable. Note that if males engage more heavily in market work while married than while single, then, given the assumed nonneutrality of education, more educated men would receive greater gains from marriage than less educated men. The predictions regarding the determinants of schooling and search time will thus differ between men and women due to the pervasiveness of the division of household time according to gender.

marriage age, since women with higher levels of schooling will be less "attractive" in the marriage market (α and T_E will be negatively correlated), given optimal investments by parents, and the less attractive search longer.

If attractiveness characteristics are indeed impounded in the error terms in equations (1), (2), and (3), the cross-equation residual correlations provide direct information on the ceteris paribus effects of attractiveness on schooling and search, since, definitionally, the errors in the consistently estimated mating equation are the attractiveness returns plus a random error. In particular, whether among observationally identical women those who are more attractive marry earlier or receive less schooling (the signs of γ_9 and ψ_7) can be inferred from the signs of the correlations between the residuals of (1) and (2) and (1) and (3), respectively.[6]

IV. Marriage Age, Spouse Selection, and Household Behavior: Fertility

The importance of characteristics unobserved by the researcher, such as attractiveness, in the determination of schooling, spouse choice, and marital search also has implications for the interpretation of empirical findings regarding the effects of "household" characteristics on consumption choices.[7] For example, the marital selection framework suggests that among the set of married couples in which the wives have the same level of education and are of the same age, households in which the husband has higher earnings will also be characterized by a wife who will have married later and/or who is relatively more attractive in the marriage market. To the extent that the life-cycle timing of household formation affects the constraints on household consumption choices and/or the attractiveness of each of the marital partners is an argument in the household utility function, associations between husband's (age-adjusted) earnings and consumption cannot be interpreted as pure income effects. Similarly, the correlation between wife's education and household consumption, for a given husband's earnings, will reflect any (life-cycle) consumption effects associated with a younger marriage age or lower marital attractiveness.

---

[6] A sufficient condition for the correspondence of the signs of the schooling and marriage-age endowment effects and the appropriate residual correlations when α is unobserved is that eqq (1)–(3) are recursive in ϵ.

[7] The analysis has implications for research on marital dissolution as well. A major hypothesis (Becker et al 1977) is that persons with lower gains from marriage have a higher probability of divorce Our results imply that among persons who marry at the same age, those with higher levels of schooling will be deficient in other respects (attractiveness), which may offset the direct gains (if any) associated with schooling Similarly, within a schooling group, persons who marry earlier will tend to have higher levels of other traits contributing positively to the gains from marriage, which may offset any negative consequences of early marriage due to the associated increase in the probability of a mismatch. The findings of Becker et al. for the United States are consistent with these selection mechanisms associated with heterogeneity, as they find no effect of educational level on the probability of marital dissolution and a U-shaped or nonexistent (for males) association between marriage age and divorce probability based on single-equation estimates. (They have other explanations for some of these results.)

With respect to one consumption choice, family size, marriage age is likely to have an important structural effect on fertility, since earlier marriage increases exposure to the risk of contraceptive failure, and thus increases contraceptive costs, and delay in marriage lowers potential fertility. As a consequence, the theory of marital selection provides an alternative explanation both for the pervasive finding in the fertility literature that mother's schooling has a negative association with fertility and for why the effect of husband's income is often negative. One conventional economic explanation for this latter finding is based on the association between the wife's value of time and education and husband's income (Willis 1973). We have shown, however, that women with higher levels of education are likely to postpone marriage in situations where there is nonmarket specialization by women in marriage and the productivity effects of schooling are market biased; thus, a negative association between female educational attainment and completed fertility may be observed even if married women with higher levels of education do not have higher values of time. Similarly, among wome [illegible]
those with a husband having highe [illegible]
married later. The negative fertility [illegible]
may offset any positive associations be [illegible]
for children associated with the usual [illegible]

The presence of heterogeneity in [illegible]
that controlling for marriage age is [illegible]
conventional demand theory relationships [illegible]
tercorrelations between unobserved marital endowments [illegible]
marriage age, and the husband's earnings potential. The "structural" equation for the fertility F of couple ij is

$$F^{ij} = b_0 + b_1 T^i_F + b_2 T^j_M + b_3 H^{ij} + b_4 A^{ij} + \epsilon^{ij}_1, \qquad (9)$$

where A^{ij} contains the combined endowments of the marital partners, including α^i. If the latter is included in the residual of a fertility equation such as (9), then OLS estimates of the household "determinants" of fertility will be biased, since, given equations (1) through (3), T^i_F, T^j_M, and H^{ij} will be correlated with the error term.

Nerlove (1974) has also suggested that the (negative) association between female educational attainment and fertility may be spurious, the consequence of heterogeneity in unobservables reflecting preference orderings. In particular, he hypothesizes that women with preferences for large families will obtain lower levels of schooling and will marry earlier. Nerlove's hypothesis thus implies that the correlations between the residuals of the fertility equation (9) and the schooling and marriage-age equations (2) and (3) will be negative, although it is not clear how or whether a woman's taste for children, given her schooling and marriage age, affects her spouse selection and thus husband's income.

However, in addition to providing signs for the residual correlations among equations (1), (2), and (3), the marital selection framework suggests that the signs of the correlations between the fertility residuals and those of the education and marriage-age equations will depend on whether the attractiveness of the marital partners is a substitute for or a complement to children in the household utility function. If, for example, attractiveness is a substitute for (or a complement to activities competitive with) child rearing—that is, the correlation between the residuals from (9) and from the mating function (1) is negative—then the marital selection framework implies that the correlations between the residuals in the fertility equation and the female schooling and marriage-age equations will be positive (more attractive women receive less schooling and marry earlier). The pattern of correlations across the structural schooling, marriage-age, mating, and fertility equations thus may provide evidence that will permit discrimination between these two competing hypotheses about the role of unobservables in fertility determination.

V. Testing the Theory: The Philippines

A. The Data

We apply the framework described by equations (1) through (9) to data from the Philippines. The Philippines provide a good laboratory for examining marital behavior and its role in fertility determination because the unique combination of features of Filipino society helps to identify the theoretical parameters of interest. First, dowry and bride prices are unimportant in the Filipino marriage market, and there is no obvious social stratification, such as a caste system, that would serve to make family background an important direct constraint on mate selection. Second, rates of labor-force participation of married women are small: in 1973, only 15.5 percent of married women were in the labor force (Rosenzweig 1976). Thus, positive assortative mating with respect to schooling should not be obscured by variations in the household division of labor, and labor-market conditions for women ought not to play an important direct role in marital matches. Third, 74 percent of those married women who were in the labor force were employed in the household. The evident compatibility of female employment with household tasks suggests that changes in employment opportunities for women may not significantly affect their gains from marriage. These characteristics imply that for Filipino women family income is not a right-hand-side variable in (1), $\beta_6 = 0$ in (1) and $\psi_3 = 0$ in (3).

The near parity of Filipino women's and men's schooling levels, despite the low market-participation rate of married women, also suggests that the schooling of women in the Philippines may be linked in important ways to the marriage market. Indeed, schooling may act as a dowry substitute in the Philippines, as most women complete their schooling prior to age 15, so that the financing of education is chiefly the responsibility of the family. The relatively young age at which schooling is completed also means that for almost all women there is a significant postschool period before marriage, as assumed in the model (the median school-to-marriage interval is 7 years). However, to the extent that schooling in the Philippines is completed prior to the full revelation of traits that are attractive in marriage, the correlation between schooling attainment, but not necessarily marriage age, and such traits may be low in the context of the Philippines.

Estimates of equations (1), (2), (3), and (9) are obtained using information from the 1973 National Demographic Survey of the Philippines, a country-wide representative sample of 8,131 households
stratified by rural and urban residence
the Philippines Population Institute
Census and Statistics Office. These
sex-specific unemployment data and
from the 1970 Census of Populati
once-married women aged 35–45 w

Table 1 defines the empirical
counterparts and provides sample descripti
or human capital, of the husband selected by each woman is measured by his predicted capitalized earnings valued at the time of marriage, based on his background, personal characteristics, and region-specific survival rate trajectories.[8] Predicted hourly earnings were

---

[8] The present value of husband's expected full income (in thousands of pesos) at time of marriage is obtained from $Y = 8.76 \int_a^\omega w(t)s(a,t)e^{-r(t-a)}dt$, where a is his age at time of marriage, $w(t)$ is predicted hourly earnings at age t, $s(a,t)$ is the probability of his survival from age a to t in the province in which the couple resides, and r (= .02) is the discount rate. The integral is multiplied by 8.76 since there are 8.76 thousand hours in a year. The survival rates are taken from the West male model life tables (Coale and Demeny 1966) corresponding to the provincial probabilities of a child's survival from birth to age 5.

TABLE 1

DESCRIPTIVE STATISTICS: MARRIED WOMEN AGED 35–44 IN PHILIPPINES (N = 1,302)

Theoretical Variable	Empirical Counterpart	Definition	Mean	SD
Endogenous.				
E_f	EDW	Years of schooling of woman	5.80	4.10
T_f	AMAR	Age of woman at marriage	20.7	4.33
H	LWH	Natural logarithm of the survival-weighted present value of the husband's earnings (in P 1,000s) at marriage	4.63	.40
F	CEB	Children ever born	6.34	2.75
Exogenous				
I	EDFA	Years of schooling of woman's father	3.21	3.46
R	SEXR	Ratio of males aged 35–44 to females aged 35–44 in province	.94	.05
I, J^m	UM	Unemployment rate for males aged 35–44 in province	4.08	2.22
P_F, J^f	UF	Unemployment rate for females aged 35–44 in province	9.57	4.15
Ω	SURV	Survival probability to age 5 (l_5) in province (thousands)	9.42	25.2
D, P_F	RURAL	Residence in rural area (dummy variable)	.54	.50

obtained from an auxiliary regression equation in which the natural logarithm of the computed hourly wages of all males (married and unmarried) was regressed against schooling level, an occupational skill index, quadratic and linear age terms, and the schooling level of the father.[9] Of the other variables, EDFA represents the income level of the wife's original home environment; the regional unemployment rates UF and UM depict market-demand levels for male and female labor;[10] the survival rate, SURV, proxies life expectancy; and the RURAL variable captures differences in population density and thus in search and schooling costs.[11] An age variable (AGEW) is also included in all equations to capture cohort effects.[12]

---

[9] Because we are using the husband's expected wage, based on a regression equation with only the husband's characteristics utilized as regressors, any effects of the wife's schooling in augmenting wealth through selection of unobserved (by the researcher) husband's characteristics affecting earnings or through the transfer of information within marriage (Benham 1974) are purged from the EDW coefficient. Thus the educational "return" estimated may understate the total (private) effect of the schooling of a woman on her husband's earnings. Note that OLS estimates of the "effects" of wife's education on husband's earnings, controlling for husband's education, will be biased downward due to heterogeneity.

[10] The cumulated past work experience of women in the 1973 sample is significantly and negatively associated with the 1970 regional female unemployment rates (Rosenzweig 1976); thus, UF and UM characterize long-term market conditions.

B. *Empirical Results: Spouse Selection, Schooling, and Search*

Table 2 displays OLS and three-stage-least-squares (3SLS) estimates of equations (1), (2), and (3), modified to reflect the Philippines setting. The 3SLS estimates of the mating function, column 2 of table 2, confirm the basic tenets of the economic theory of marriage—a delay in marriage by women (longer search duration) is associated with a husband with greater expected earnings power while, given search time, the higher the level of the wife's schooling the more likely she is to have married a husband with high earnings potential. The 3SLS point estimates suggest that the gross "returns" to female schooling in terms of the husband's expected present value of income, the percentage increase in expected wealth due to marriage choice resulting from a 1-year increase in the wife's schooling, is 5 percent. An additional year of delay in marriage increases the expected wealth of the husband by 16 percent at the time of marriage.[13] Moreover, as indicated by the theory, for given daughter (wife) traits and behavior, the higher the number of males to females in the area and the lower the population dispersion the greater the earnings of the marital partner.

While none of the statistically significant OLS mating function coefficients shift importantly when the 3SLS estimation technique

---

[11] The estimate of SURV is derived from
ever born and children surviving to women
described in Trussell (1975) Use of this index
not the shape of the life-table survival probabili
from birth to age 5 (l_5), an assumption employe
Coale-Demeny (1966) life tables Thus, if mor
the same regional model life table, SURV is p

[12] Because no information on age at marriage
vided for never-married women, who are excluded fro
still be affected to some degree by selectivity bias We have
age floor in obtaining our sample, as census data indicate that almost 85 percent of all women had been married by that age in 1970 (the proportion ever married by age 15-49 is 91). *The coefficients of AGEW in the marriage-age equation* will thus reflect both cohort and selectivity effects to some extent

[13] Auxiliary structural estimates indicate that higher levels of female schooling and female marriage age are associated with the marriage of a more educated husband However, while a delay in marriage results, ceteris paribus, in a younger husband, the higher is female schooling for given female marriage age, the older is the husband The relative marital returns to female schooling and marriage age thus depend on the age trajectories of earnings and survival probabilities and the discount rate The two latter characteristics are given by the data and are used in computing the wealth variable Experimentation with the discount rate indicated that there was an inverse relationship between the estimated magnitude of the (positive) 3SLS effect of AMAR on LWH and the level of r up to r = 10

TABLE 2

STRUCTURAL EQUATION ESTIMATES: SCHOOLING, AGE AT MARRIAGE, AND THE MATING FUNCTION
(Standard Errors in Parentheses)

	LWH		EDW		AMAR	
VARIABLE	OLS	3SLS	OLS	3SLS	OLS	3SLS
EDW*	056	.050	...	. .	.304	.215
	(.002)	(.024)			(031)	(.061)
AMAR*	−.003	.162	..	.	..	...
	(002)	(089)				
EDFA	..	. .	566	.565	.	..
			(.028)	(028)		
UF	.	. .	064	.058	..	...
			(.029)	(.029)		
SEXR	−.173	1 43	−.967	−.806	−9.30	−8.59
	(129)	(.948)	(1 86)	(1 86)	(2.42)	(2 37)
SURV	...	...	.009	.010	.001	008
			(004)	(004)	(.005)	(.002)
UM	− 018	−.024	−.098	−.091	.028	.031
	(.003)	(.012)	(.060)	(.060)	(.063)	(063)
RURAL	− 338	−.238	−2.36	−2.32	.350	137
	(015)	(.062)	(.231)	(.230)	(.276)	(329)
AGEW	− 002	−.018	−.029	−.029	.110	.106
	(.002)	(012)	(032)	(032)	(042)	(.042)
Constant	4.88	555	− 953	−2 076	22.45	15 66
	(171)	(2 21)	(473)	(4.61)	(6.09)	(3 74)
R^2	690	.	.408	..	099	.
Wu†	..	610.5	.	.	..	8.82

* Endogenous variables
† From 2SLS estimates

is applied, use of ordinary least squares appears to understate significantly, as predicted, the effect of marital search by women in attracting a "better" mate—indeed, while the 3SLS AMAR coefficient is statistically significant (5 percent level, one-tail test) and positive, the corresponding OLS parameter is insignificant and negative. Moreover, application of the test suggested by Durbin (1954) and Wu (1973) results in rejection of the hypothesis that schooling and marriage age are uncorrelated with the mating function residuals; the computed Wu statistic is 610.5, well above the critical F-value for 2,1200 degrees of freedom. The OLS bias in the AMAR coefficient is thus consistent with the presence of negative "selectivity" associated with the length of single status. This bias is also evidently imparted to the coefficient of SEXR, which is only positive and statistically significant when 3SLS is employed.

The parameter estimates in the schooling and marriage equations also tend to conform to the hypotheses about the determinants of educational investment and marital search. The schooling equation estimates indicate that women with more educated and presumably higher-income fathers and in areas where life expectancy is relatively high obtain higher levels of schooling, with the coefficients of SURV and EDFA statistically significant at the .05 level. The sign of the statistically significant female unemployment rate coefficient in the education equation is also "correct"—higher rates of female unemployment, and thus lower opportunity costs of schooling, are associated with greater levels of female schooling attainment. Schooling attainment is also lower in rural areas, where costs associated with travel time to school are presumably higher than in urban areas. While the sex-ratio coefficient in the schooling equation also displays the predicted sign—where males are relatively more prevalent schooling investment in daughters is less—the coefficient is not statistically significant by conventional standards.

In the 3SLS marriage-age equation, the coefficient signs for schooling, SURV, male unemployment rate, population density (RURAL), and the sex ratio are in accord with the marital search framework—increases in the schooling level of women, increases in survival prospects, poorer labor-market prospects for men, and greater population dispersion raise the age at which women marry, while increases in the relative number of males significantly lower age at marriage Moreover, as predicted, the OLS estimate of the coefficient of EDW is upwardly biased (by 40 percent) due to heterogeneity and the endogenous response of schooling to exogenously give [illegible] the marriage market, although this diff [illegible] cally significant. Indeed, the Wu statisti [illegible] the .05 level, unlike that for the mating [illegible] the correlation between EDW and the [illegible] due, as noted, to the low average ag [illegible] Philippines.

The pattern of residual correlations am [illegible] schooling, and marriage-age equations, reporte [illegible] to support the hypothesis that personal char [illegible] mates but unrecorded in the data influence marital search and, to a lesser extent, schooling. Since the residuals from the LWH mating equation contain the marital returns to unobserved female endowments, as noted, the negative correlations between these residuals and

---

[11] As stressed by Nakamura and Nakamura (1983), the hypothesis tested concerns the statistical significance of the correlation between a suspected endogenous variable X and the structural equation residual, not the significance of the change (= bias), if any in the coefficient of X resulting from the use or nonuse of instruments. Moreover, their results suggest that the power of the Wu test is low, so that for large samples the null hypothesis of zero correlation will almost always be rejected for very small nonzero correlations.

TABLE 3

RESIDUAL CORRELATION MATRIX

Variable	EDW	AMAR	LWH
EDW	...	.073	–.249
AMAR	. .	.	– 744
CEB	149	.428	–.453

those from the education and AMAR equations are supportive of the spouse selection framework, which suggested that more attractive daughters would receive lower levels of schooling and would remain unmarried least long.[15] As expected, however, the correlation of the EDW residuals with the mating function residuals is much less strong than that between the residuals of the LWH and AMAR equations.

Finally, to test the overidentification restriction that parental background has no direct effect on the quality of the spouse attracted by the daughter, EDFA was added to the mating and marriage-age equations. However, we could not reject the hypothesis that family background has no influence on spouse's earnings characteristics for given daughter's traits and search, because EDFA did not attain statistical significance in either equation. The influence of a woman's parents in the Philippines on her age at marriage and the quality of her mate is evidently only indirect, via parental financing and/or determination of schooling investment.

C. *Empirical Results: Spouse Selection, Tastes, and Fertility*

Table 4 reports estimates of variants of the structural fertility equation (9). The first specification, reported in column 1, conforms to the "conventional" household fertility demand equation in which the wife's schooling, husband's (permanent) income, survival rates, location, and age are regressors. The specifications in columns 2 and 3 both add the wife's age at marriage. The estimates in column 2, like those in column 1, are obtained using OLS, while the estimates in column 3 are obtained using 3SLS as part of the equation system including the mating function and the education and marriage-age equations. The estimates in the third column thus take into account

---

[15] Taylor and Glenn (1976) find, based on a U.S. probability sample, that for most socioeconomic background groups, physically more attractive women, for given schooling, and women with higher levels of schooling, for given physical attractiveness, marry men with higher levels of socioeconomic attainment. These associations, however, are weaker for daughters of farmers. The authors hypothesize that this anomaly arises because the more attractive women from this socioeconomic origin class tend to marry earlier than their less attractive siblings, but they cannot test the hypothesis because of the absence of data on age at marriage.

TABLE 4

DETERMINANTS OF CHILDREN EVER BORN: ESTIMATED FERTILITY EFFECTS OF FEMALE SCHOOLING, AGE AT MARRIAGE, AND HUSBAND'S EARNINGS
(Standard Errors in Parentheses)

Variable/ Estimation Technique	OLS	OLS	3SLS
ED*	– 177	–.067	- 042
	(026)	(.024)	(212)
LWH*	.476	091	150
	(327)	(289)	(2 23)
AMAR*		– 296	– 551
		(.016)	(.167)
SURV	–.006	– 003	– 001
	(003)	(.003)	(005)
RURAL	361	281	162
	(193)	(171)	(484)
AGE	.123	155	.179
	(024)	(021)	(033)
Constant	5 65	9.44	11.35
	(3.18)	(2 82)	(7 67)
R^2	096	295	

* Endogenous variables

marital behavior and population heterogeneity. Note that identification of the fertility equation is achieved by again making use of the observation that few married Filipino women are employed outside the home, so that changes in female emp
do not alter the relative costs of housel

The OLS estimates of the wife's ed
effects on children ever born in the fir
with conventional results, with both va
cantly associated with the fertility meas
cant, negative female schooling effect, in
the hypothesis that women with higher
fewer children. The 3SLS estimates in
that the schooling and survival effects on fe
marital selection.[16] Neither female education, husband's
the province-level survival rate remains a significant determinant of children ever born when both heterogeneity and marriage age are taken into account.[17] Indeed, the estimates indicate that a 1-year in-

---

[16] The 3SLS results are similar to Harman's (1970) findings based on data from the 1968 National Demographic Survey of the Philippines Harman treated marriage age, but not female schooling or husband's income, as an endogenous variable in a simultaneous equations system

[17] Of course, ED and SURV affect the choice of age at marriage and thus LWH Reduced-form estimates suggest that fertility is significantly higher in rural areas and where males of marriage age are relatively abundant; women from households in which the father has a higher level of schooling bear significantly fewer children

crease in the schooling of a woman, for a given marriage age, would lower children ever born by only 0.04 children, while a delay in marriage by 1 year would lower fertility by more than one-half child, an elasticity of − 1.8.

The OLS estimates in column 2 suggest that the large reduction in the female education coefficient is due mainly to the strong correlation between female schooling and marriage age, which itself has a strong negative structural effect on fertility. The operation of selectivity associated with heterogeneity in traits valued by marital partners is also evident, however. The test statistic for the correlation of the three endogenous variables EDW, LWH, and AMAR with the fertility residual is 136.4, well above the critical F-value for 3,1200 degrees of freedom; indeed, use of the appropriate 3SLS estimation procedure, while not reducing the education coefficient significantly, results in an increase in the absolute value of the statistically significant, OLS age-at-marriage coefficient by 87 percent. The correlation between the residuals from the structural LWH equation, containing the returns to the wife's unobserved endowments, and the structural fertility equation residuals, reported in table 3, moreover, is large and negative. Among observationally identical couples, those with a wife who has characteristics more valued in the marriage market evidently substitute away from children in their consumption allocations.

The negative correlation between marital attractiveness and fertility implies, as was noted, that if the utility-yielding traits of the marital partners dominate the error term in the structural fertility equation, the fertility residuals will be positively correlated with those from the marriage-age equation. They will also be correlated, to a lesser extent (in the Philippines), with those from the schooling attainment equation due to marital selection. This is confirmed in table 3—the residual AMAR-LWH correlation is .43, that between EDW and LWH is .15. These latter correlation patterns are, moreover, inconsistent with Nerlove's hypothesis that women with above-average preferences for children obtain less schooling and marry earlier, which would imply that these correlations would be negative. Variations in marriage age and in marital attractiveness appear to account for almost all of the systematic variation in fertility among Filipino women.

VI. Conclusions

In this paper we have shown how the existence of linkages between human capital accumulation, spouse selection, and exogenous traits known to marriage-market participants but unobserved by the researcher may obscure some of the predictions of the economic theory of marriage based on estimates predicated on the exogeneity of all personal attributes. The selectivity associated with marital sorting was also shown to importantly affect the interpretation and estimates of how household characteristics such as income and education are related to one important choice of households, family size.

The empirical findings from the Philippines provided support for the hypothesis that schooling, marital search, and spouse selection are endogenous variables influenced directly or indirectly by the total resource of parents, endowed traits of offspring, the costs of schooling, and marriage-market conditions. Instrumental estimation techniques, used to eliminate bias due to the presence of heterogeneity in marital endowments, but not OLS, provided confirmation of two major but hitherto untested tenets of the economic theory of marriage—namely, that there are payoffs to spouse search and positive assortative mating with respect to schooling, which is evidently valuable even when labor-force participation is low. Our results also indicated, however, that, while additional schooling attracts a higher-value spouse, it lowers the gains from marrying. Consequently, women with more schooling and less attractive women tend to marry later than other women. There was some evidence suggesting that less attractive women also tend to receive more schooling.

Finally, the results rejected the hypotheses that more educated women in the Philippines have lower fertility because their value of time is higher and that more educated Filipino women have lower preferences for children. Support was provided instead for the hypothesis that the observed female edu
the Philippines reflects the optimal s
agents heterogeneous in marriage-m
for children in household consumptio
the relationships between household
household consumption behavior wou

References

Anderson, Kathryn H., and Hill, M. Anne. "Determinants of Age at Marriage in Malaysia and Japan." Econ. Growth Center Discussion Paper no. 351. New Haven, Conn.: Yale Univ., June 1980.

Becker, Gary S. "A Theory of Marriage: Part I." *J.P.E.* 81 (July/August 1973): 813–46.

———. "A Theory of Marriage: Part II." *J.P.E.* 82, no. 2, pt. 2 (March/April 1974): S11–S26.

Becker, Gary S.; Landes, Elisabeth M.; and Michael, Robert T. "An Economic Analysis of Marital Instability." *J.P.E.* 85 (December 1977): 1141–88.

Becker, Gary S., and Tomes, Nigel. "An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility." *J.P.E.* 87 (December 1979): 1153–89.

Behrman, Jere, and Wolfe, Barbara. "Who Marries Whom? And How It

Affects the Return to Schooling." Mimeographed. Philadelphia: Univ. Pennsylvania, April 1983.

Benham, Lee. "Benefits of Women's Education within Marriage." *J.P.E.* 82, no. 2, pt. 2 (March/April 1974): S57–S71.

Coale, Ansley J., and Demeny, Paul. *Methods of Estimating Fertility and Mortality from Censuses of Population.* Princeton, N.J.: Princeton Univ. Press, 1966.

Durbin, James. "Errors in Variables." *Rev. Internat. Statis. Inst.* 22 (1954): 23–32.

Elder, Glen H., Jr. "Appearance and Education in Marriage Mobility." *American Soc. Rev.* 34 (August 1969): 519–33.

Frieden, Alan. "The United States Marriage Market." *J.P.E.* 82, no. 2, pt. 2 (March/April 1974): S34–S53.

Harman, Alvin J. *Fertility and Economic Behavior of Families in the Philippines.* RM-6385-AID. Santa Monica, Calif.: Rand Corp., 1970.

Keeley, Michael C. "The Economics of Family Formation." *Econ. Inquiry* 15 (April 1977): 238–50.

Nakamura, Alice, and Nakamura, Masao. "On the Performance of Tests by Wu and by Hausman for Detecting the Ordinary Least Squares Bias Problem." Mimeographed. Toronto: Univ. Toronto, March 1983.

National Census and Statistics Office. *Philippines 1970 Census of Population and Housing.* Manila, 1974.

Nerlove, Marc. "Household and Economy: Toward a New Theory of Population and Economic Growth." *J.P.E.* 82, no. 2, pt. 2 (March/April 1974): S200–S218.

Rosenzweig, Mark R. "Female Work Experience, Employment Status, and Birth Expectations: Sequential Decision-Making in the Philippines." *Demography* 13 (August 1976): 339–56.

Taylor, Patricia A., and Glenn, Norval D. "The Utility of Education and Attractiveness for Females' Status Attainment through Marriage." *American Soc. Rev.* 41 (June 1976): 484–98.

Trussell, T. James. "A Re-Estimation of the Multiplying Factors for the Brass Technique for Determining Childhood Survivorship Rates." *Population Studies* 29 (March 1975): 97–107.

Willis, Robert J. "A New Approach to the Economic Theory of Fertility Behavior." *J.P.E.* 81, no. 2, pt. 2 (March/April 1973): S14–S64.

Wu, De-Min. "Alternative Tests of Independence between Stochastic Regressors and Disturbances." *Econometrica* 41 (July 1973): 733–50.

Entry, Industry Growth, and the Microdynamics of Industry Supply

John C. Hause

State University of New York at Stony Brook

Gunnar Du Rietz

Svenska Arbetsgivareforeningen

Entry is widely discussed but rarely subjected to empirical study. This study develops a competitive theory of entry, with primary focus on the relationship between entry and industry growth. The main ingredients are adjustment costs to firms already in the industry and the distribution of fixed entry costs to potential entrants. The theory suggests sufficient conditions under which the entry rate is an increasing, convex function of the industry growth rate. A regression model correcting for severe heteroscedasticity is applied to data from Swedish manufacturing industries. Th
with the theoretical prediction for gro
expected to influence entry significan

I. Introduction

Recognition of the central role of entry i
least as old as Adam.[1] Oddly enough, the subsequ

Hause's work on this study has received support at various stages from the Industriens Utredningsinstitut, from a visiting fellowship at the Hoover Institution, and from NSF grant SES-7927217. He is indebted to George J. Stigler for initial interest in this topic, to Ken Hendricks, and to an unknown referee for constructive comments. H. C. Yeh carried out most of the calculations reported in Sec. IV of the study. Only the authors are liable for the analysis and conclusions that follow.

[1] "When by an increase in the effectual demand, the market price of some particular commodity happens to rise a good deal above the natural price, those who employ their stocks in supplying that market are generally careful to conceal this change. If it was

[*Journal of Political Economy*, 1984, vol. 92, no. 4]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9204-0003$01.50

empirical elaboration of entry has been like *Moby Dick* without the whale. The main line of development has followed John Bates Clark's doctrine of "potential competition" (1887), recently resurrected and dressed in formal attire by Baumol, Panzar, and Willig (1982). Clark was concerned with understanding how competitive forces would operate if the atomistic structural competition that characterized agriculture and precorporate manufacturing were replaced by industrial markets dominated by a few large firms. He concluded that even a formal monopolist may well charge a "normal" price, reflecting only his costs of production, if there are potential entrants who could and would appear if price rose above this level because of monopolistic restriction of output. Clark believed that recognition of this potential competition was an important constraint on the behavior of formal monopolists and trusts. Over time, he became increasingly skeptical of the adequacy of this control of monopoly power, unless policies were adopted to prevent predatory behavior that might otherwise discourage entry.[2]

The past 25 years have seen the development of a related literature on "barriers to entry," stemming from papers by Bain (1949) and Modigliani (1958). This work attempted to characterize theoretically conditions under which significant long-run departures from Clark's "normal price" might occur. And it attempted to measure empirically the size of such departures, in terms of rates of return or price-cost differentials. It would seem natural for a much more systematic investigation of actual entry to complement this work. Although several empirical studies of actual entry have been carried out, very limited attention has been given to this topic.[3]

There are several reasons why realized entry deserves more attention. (1) Economic theory says almost nothing about where the output comes from when an industry expands. Under what conditions will it come almost exclusively from the firms already in the industry? When

---

commonly known, then great profit would tempt so many new rivals to employ their stocks in the same way, that, the effectual demand being fully supplied, the market price would soon be reduced to the natural price, and perhaps for some time even below it. If the market is at a great distance from the residence of those who supply it, they may sometimes be able to keep the secret for several years together, and may so long enjoy their extraordinary profits without any new rivals. Secrets of this kind, however, it must be acknowledged, can seldom be long kept; and the extraordinary profit can last very little longer than they are kept" (Smith [1776] 1937, p. 60).

[2] See Stigler (1968) for several interesting comments on potential competition

[3] Previous cross-sectional studies of entry include Hause (1962), Mansfield (1962), Wedervang (1965), McGuckin (1972), Orr (1974), Duetsch (1975), Du Rietz (1975, 1980), and Gorecki (1975). This study is a major refinement and extension of work reported in Du Rietz (1975, 1980).

will entry play an important role in such expansion? Our inability to answer such questions adequately indicates a major lacuna in the theory of competitive supply. (2) The importance of potential competition in generating competitive outcomes depends on the sensitivity of entry to changes in the returns to or costs of entry. Past attempts to measure barriers to entry have not utilized information on differential entry rates to address this issue. Some of the variables determining entry rates can be interpreted as costs. The sensitivity of entry to variations in such costs is useful evidence on the potential importance of potential competition.

This study is focused primarily on entry by new firms. In Section II we provide a quantitative summary of this entry for a 15-year period (1954–68) for most of the plastics, primary metals, and engineering manufacturing industries in Sweden. We also consider entry by established firms, which diversify by building a plant in a new industry (for them). In Section III, we develop a simple theory of the relationship between entry and the growth rate of a competitive industry. It is important to understand the mechanism of entry in a competitive context before claiming a serious understanding of the role of strategic behavior to influence entry in markets with monopoly power. Several other variables that affect entry rates are discussed briefly. Section IV presents some econometric evidence on determinants of entry, with particular emphasis on the industry growth rate. This evidence suggests that entry is an increasing, convex function of industry growth rates, a result consistent with the theoretical argument in Section III. This empirical work mak[illegible] developed extension of the usual regress[illegible] heteroscedasticity, an important statis[illegible] section industry studies.

II. Market Shares of New and Divers[illegible] Entrants

A natural way to characterize the direct impact of entry over a period of time is by the end-of-period market share of the entrants. Market shares are measured by employment, an imperfect but adequate surrogate for product market shares. Table 1 reports the entrant shares for 39 Swedish manufacturing industries for the 15-year period ending in 1968. The share of new firms ranges from 0 to 26 percent, while the share of firms that have diversified into new industries ranges from 0 to 20.7 percent. We denote these two types of entry by E_n and E_d in the following discussion. The share of entry from both

TABLE 1

DIRECT IMPACT OF ENTRY ON 39 SWEDISH MANUFACTURING INDUSTRIES

INDUSTRY	ENTRY AS % OF TOTAL 1968 EMPLOYMENT		TOTAL EMPLOYMENT
	New Firm	Diversification	
351310 Basic plastics	3.7	1.5	1,892
351320 Semimanufactured plastic goods	7.8	20.7	5,164
356000 Manufactured plastic goods	21.0	9.3	8,052
371010 Iron and steel mills	0	.3	42,000
371030 Iron and steel foundries	1.5	1.4	4,733
372030 Nonferrous metals	2.3	0	7,783
372040 Nonferrous foundries	6.4	1.6	1,969
381100 Cutlery, hand tools, and hardware	12.3	1.2	6,756
381200 Metal furniture	15.1	.8	4,139
381300 Structural metal products	24.1	3.7	20,095
381910 Metal containers	19.3	2.3	2,839
381920 Wire cloth, wire, and cable	1.6	4.0	4,675
381930 Nails, bolts, and nuts	1.1	2.3	4,652
381940 Other metal products for construction	3.4	3.7	11,486
381950 Household metalware	3.3	0	4,212
381990 Other metal products	15.4	2.9	18,818

382200 Agricultural machinery and equipment	4.1	1.3	8,585
382310 Metalworking machinery	2.6	1.3	7,577
382320 Woodworking machinery	2.7	0	2,333
382410 Pulp and paper mill machinery	6.7	0	2,955
382420 Construction and mining machinery	6.5	0	6,369
382490 Industrial machinery N.E.C.	9.7	.5	13,412
382590 Other office and accounting machinery	3.0	5.2	10,673
382991 Lifting and hoisting machinery	6.1	5.2	11,344
382992 Liquid pump manufacturing	8.5	0	3,329
382993 General purpose machine parts	1.1	.9	9,182
382999 Other machinery and equipment	4.1	1.0	26,073
383100 Electrical industrial machinery and apparatus	2.5	1	15,160
383200 Radio, TV, and communication equipment and apparatus	1.3	.3	28,230
383300 Electrical appliances and housewares	4.3	.02	11,535
383910 Insulated wires and cables	3.9	0	4,102
383920 Storage batteries and accumulators	.2	5.4	2,296
383930 Light bulb and fluorescent lamps	2.9	0	1,621
383990 Other electrical equipment	11.4	1.6	5,114
384110 Ship building and repairing	.6	0	25,833
384120 Boat building and repairing	26.0	4.7	1,834
384310 Motor vehicles and chassis	.4	0	15,250
384320 Motor vehicle engines, parts	7.1	1.3	19,676
384400 Motorcycles and bicycles	1.3	0	1,462
Total	5.8	1.7	383,210

sources ranges from 0.3 to 30.7 percent. At a more aggregate level, E_n amounts to 5.8 percent for the 39 industries, while E_d is 1.7 percent.[1]

If entry is measured instead by the fraction of existing firms in 1968 that entered the industry as new or diversifying firms over the preceding 15 years, E_n ranges from 0 to 46 percent for the 39 industries. E_d entry measured the same way ranges from 0 to 20 percent.

The first impressions from these data are that entry usually accounts for a rather modest share of an industry's activity, even after a 15-year period, although there is a large variance across industries. Entrants do account for 20 percent or more of industry employment for five industries in our sample, although these industries are rather small ones. Entry measured by relative numbers of firms exaggerates the direct impact of entry by new firms, since new firms tend to be much smaller than established firms.

Most economists would probably agree that high rates of entry can be regarded as prima facie evidence that long-term monopoly is not sustainable, even if attainable. Based on the 7.5 percent aggregate employment share of E_n and E_d, some observers might conclude that entry is not a very important mechanism for enforcing competitive behavior in industries. We disagree with this inference. The potential competition literature makes the important point that a high observed entry rate is not a necessary condition for obtaining approximately competitive levels of output and price even with substantial industrial concentration. The sensitivity of entry to changes in the costs or returns to entry is more important than the level of entry as an indicator of the importance of potential competition. Some evidence on this sensitivity is provided in Section IV.

III. The Relationship between Entry and Industry Growth Theory

This section develops a simple competitive theory of the qualitative relationship between entry and industry growth. It concludes with a brief, informal discussion of several other important determinants of entry and defines the empirical counterparts to the theoretical variables.

Our theory has two ingredients, (1) costs of adjustment to existing firms in altering their capital stock and (2) the distribution of fixed costs of entry among potential entrants. The discussion initially omits the possibility of firm exit from the industry. The analysis leads to sufficient conditions under which entry is an increasing, convex function of the industry growth rate.

---

[1] The entry data are from establishment statistics. Each establishment is assigned to one industry even for multiproduct establishments. Data were also collected on entry when an establishment (or entire firm) is reclassified into a new industry, but are not analyzed in this paper. Such "entry" is small relative to new firm and diversification entry and contains large measurement errors. See Du Rietz (1975, 1980) for further information.

Consider an industry producing a single homogeneous product X with two factors of production, capital K and labor L. All firms are competitive (price takers) in the product and factor markets. Firms have rational expectations about future values of the product price $P(s)$ and factor prices $G(s)$ and $W(s)$ (for gross investment and wages), where s is any future point in time. These expectations are consistent with the firm decisions to maximize discounted net receipts. The discount rate r and the depreciation rate of capital m are constants.

Firms have production functions that are identical and homogeneous of degree one,

$$X(K, L, I) = F(L, K) - KC\left(\frac{I}{K}\right), \tag{1}$$

where gross investment $I = (dK/dt) + mK$, the sum of net investment, dK/dt, and the constant depreciation rate of the capital stock, mK. This specification of adjustment costs as part of the production function follows Lucas (1967). F is twice continuously differentiable and satisfies the usual production function restrictions that both factors of production have positive marginal products ($F_K > 0$, $F_L > 0$) and generate the usual convex isoquants. The adjustment cost is additive, and it is assumed that C is twice continuously differentiable and possesses a piecewise continuous third-order deriv
restricted by the condition $I \geqq 0$, so that ne
(negative) bound equal to the depreciation r
$C(0) = 0, C' > 0, C'' > 0$, and $C''' > 0$. These
marginal adjustment cost function is positive
The convexity assumption is discussed below
sured by its capital stock, the growth rate of
Therefore, the argument of the C function can
$+ m)$ when convenient. Since the depreciation rate of the capital stock is constant and exogenous, optimal choice of the gross investment rate $I(s)$ by the firm is equivalent to optimal choice of its growth rate $g(s)$.

A firm's objective is maximization of the discounted net cash flow from producing X (or exiting from the industry, if that increases its net worth). The formal problem is to choose $I(s)$ (or $g[s]$) and $L(s)$ for $t \leqq s \leqq \infty$ to maximize

$$\int_t^\infty \left\{P\left[F(K, L) - KC\left(\frac{I}{K}\right)\right] - WL - GI\right\} \exp(-rs)ds \tag{2}$$

Appendix A derives the marginal conditions for an interior optimum:

$$PF_L = W \tag{3a}$$

and

$$\int_t^{\infty} PX_K(s) \exp\left[-(r + m)(s - t)\right]ds = PC' + G. \tag{3b}$$

Equation (3a) states that the value of the marginal product of the costlessly variable factor must be equal to its price for a competitive firm. Equation (3b) equates the discounted value of a marginal unit of gross investment at the current point in time t to the current cost of an additional unit of gross investment. The latter is the sum of two components, (1) the current market price of X times the reduction of current output because of the marginal adjustment cost and (2) the market price of a unit of gross investment, G. We follow Mussa (1977) in calling the left-hand side of (3b) $P_K(t)$, the shadow (demand) price of capital, while we call the right-hand side the marginal cost of capital. From the preceding assumptions about C, it immediately follows that the marginal cost of capital is an increasing, convex function of gross investment I (equivalently, of the firm growth rate g). Since the firms in this model are scale replicas of each other (depending on when they entered the industry and on entry size) and the firms have identical rational price expectations, all firms in the industry have the same $P_K(t)$ and the same optimal growth rate $g^*(t)$. Hence, the marginal cost of capital for the "old" firms already in the industry is a function of the equilibrium growth rate of these firms, $j_o(g_o)$.

Before analyzing entry, we pause to consider more closely the assumption that the marginal adjustment cost function is increasing and convex. To our knowledge, there is no discussion in the literature whether it is convex or concave, although a recent diagram by Mussa (1977, p. 166) implicitly assumes that it is convex.[5]

In the analysis of entry, the shape of the marginal adjustment cost function plays an important role in determining the behavior of entry as industry growth becomes large. The standard assumptions of production theory are not restrictive enough to answer this question, although the hypothesis of positive, increasing, and weakly (if not strictly) convex adjustment costs (or costs of financial capital) seems

[5] The large literature on adjustment costs of firm growth (see, e.g., Holt et al. 1960, Gould 1968, Treadway 1969, 1970; Soderstrom 1977; Brock and Scheinkman 1978) generally assumes that the marginal adjustment cost is an increasing function of gross or net investment. See Rothschild (1971) and Nickell (1978, pp. 35–40, 256–71) for a more critical view. None of the published literature appears to consider the convexity or concavity of the marginal adjustment cost function.

plausible to us, at least for high firm growth rates. There are two dimensions to be distinguished in firm capacity expansion, (1) the relative size of the increment of capacity (at completion) to initial capacity and (2) the length of time over which the expansion takes place. When the first component is held constant, casual observation provides support for the adage "haste makes waste." New oil refineries, paper mills, and steel plants are examples of complex construction projects that are commonly staged over substantial time. The decision of firms to assume the interest costs accumulated in such undertakings before output is available can only be explained by the expectation that the cost would be still greater if the firm attempted to reduce the construction period.[6] Experience with crash programs to expand production capacity or to achieve some technological feat demonstrates costs that again seem to support the view of accelerating costs as the project construction time is shortened. Simple queuing models can be constructed yielding the conclusion in a more formal way. Thus we consider weak convexity of the marginal cost function a reasonable hypothesis for high firm growth rates.

We next consider potential entrants and formal conditions that must be satisfied to induce entry to the industry. There is a pool of potential entrants to the industry, consisting of new firms that would be formed if profitable and of existing firms that would enter if entry would increase their present value. These potential entrants have the same price expectations and production function (1) as firms in the industry. Entrants are distinguished from other firms by fixed costs of entry, which are a function of the capital stock with
enters, K_i. $Z_i(K_i)$ denotes these fixed costs for
$Z_i(0) = 0$, $Z_i' > 0$, and $Z_i'' > 0$, at least in th
optimum entry size. The ith potential firm c
stock K_i^* and planned growth rate that maxi
tween equation (2) and $Z_i(K_i)$ if the differen

The Z_i for potential entrants captures both
costs of entry associated with differences in sp
differences in learning costs of becoming an effic
industry. The condition for profitable entry is given by a critical price of capital for the ith firm, $P_{K_i}^*$, such that entry of the firm occurs if the calculated shadow price is greater than or equal to this value. In this circumstance, the firm brings in K_i^* units of capital. It seems plausible that the density function of capital that will be brought into the industry by potential firms is a unimodal function of the price of capital

---

[6] Strictly speaking, the production function in (1) does not properly capture investment processes that go on for more than one time period before the new capacity is completed and able to deliver output

and, furthermore, that only rarely (if ever) will the shadow price of capital be so high that the mode of the entrant density function of capital is observed.[7] To obtain conclusions about the effect of industry growth rates on entry, we use discrete time language and consider the density function of *industry growth from entrants* ($g_e = K_{eT}/K_0$) as a function of P_K, the shadow price of capital. Here K_{eT} is the amount of capital introduced by entrants in the time interval from zero to T; K_0 is the total industry stock of capital at the beginning of the period. We denote this density function by $\phi(P_K)$. Presumably $\phi(P_K) = 0$ for $P_K \leqq G$, the market price of gross investment. Then the industry growth from entry is

$$g_e = h_e(P_K) = \int_G^{P_K} \phi(x)dx. \tag{4}$$

Assuming that $\phi(P_K)$ is unimodal and that the P_K's that occur are less than the modal P_K, it immediately follows that g_e is an increasing convex function of P_K and that the inverse function of g_e, $j_e(g_e)$, is an increasing, concave function of g_e (see fig. 1).

Since the shadow price of capital, $P_K(t)$, is the same function of time for all old and potential firms in this model, the horizontal sum of the graphs of $j_o(g_o)$ and $j_e(g_e)$ for the old and new firms is the graph of the equilibrium shadow price of capital for the firms as a function of the industry growth rate $g_x = g_o + g_e$. Hence both g_x and g_e are functions of the parameter P_K. From the last panel of figure 1, g_e is an increasing, convex function of g_x; that is, if $g_e = u(g_x)$, then $u' > 0$ and $u'' > 0$. The last panel shows two equal-sized increments of industry growth on the horizontal axis at a low and a high industry growth rate. It projects these two increments back onto the entry curve and down to the entry axis, showing the size of the entry component in these two increments of industry growth. Given the curvature of the entry curve and of the growth curve for old firms, it is obvious that the entry component of the industry growth increment must be larger for the growth increment at the higher growth rate. This establishes the convexity of g_e as a function of g_x. The conditions that $j_o(g_o)$ be an increasing, strictly convex function and that $j_e(g_e)$ be an increasing,

---

[7] This conjecture stems from the following intuitive line of argument. We expect a positive correlation between P_K and the transitory profits required to offset the fixed entry cost, Z_i, of the ith potential entrant. The Z_i's reflect the learning costs and quasi rents of specialized resources owned by the potential entrant. Historical episodes of exploration for gold and oil and wartime expansion of output by entrants illustrate the volume of resources that entrants will bring into an industry when transitory profits are high. But we doubt whether the shadow price of capital in a well-defined industry ever becomes so high that most entrepreneurs in the economy seriously contemplate entry. We expect that the mode occurs at a shadow price of capital substantially above the usual range of values taken on by this shadow price.

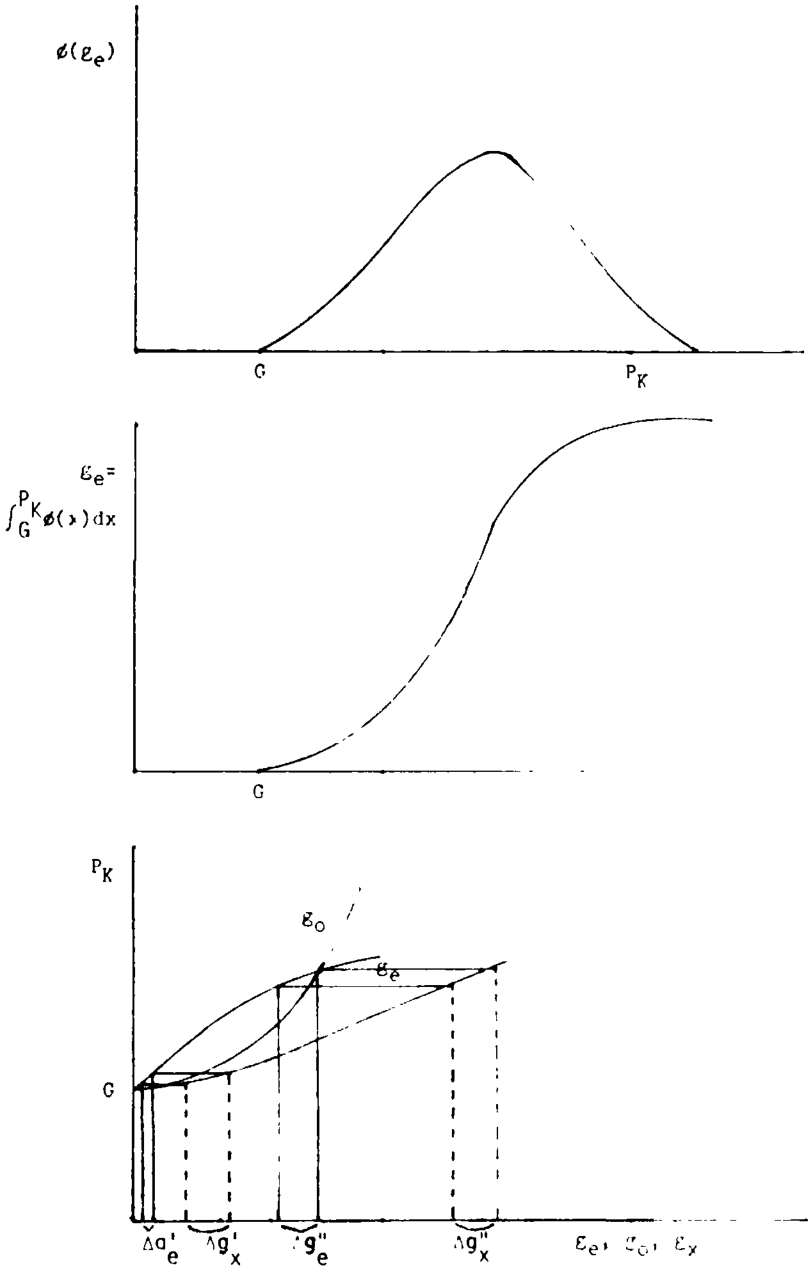


FIG. 1.—Density function of entry, entry function, and old firm growth as functions of P_K (shadow price of capital).

strictly concave function are obviously sufficient but not necessary to assure that g_e is an increasing, convex function of g_x.[8]

A slight modification of the model allows for firm exit from the industry. Each firm can sell its assets and dissolve or can exit from the industry and enter another if its present value from remaining in the industry (given by eq. [2]) falls below some critical value. An analogous argument about the distribution function of capital that will be withdrawn from the industry if the shadow price of capital falls below a critical level preserves sufficient conditions for the conclusion that g_e is an increasing, convex function of g_x.[9]

The assumption of identical cost structure for all firms in the industry (1) is convenient for theoretical analysis, since P_K is then a common parameter for all firms. If the cost-of-adjustment function differs between firms, it would be necessary to relate the individual adjustment costs to the disturbance generating a change in the industry's capital stock.[10]

The preceding analysis of the entry-industry growth relationship tacitly assumes a positive industry growth rate. If the industry is contracting, one still expects some new firm entry will occur, replacing some old firms that die or exit to other industries. This "turnover" entry is discussed in Hause (1962). To allow for its presence in the empirical analysis of Section IV, we define two industry growth rate variables. We use x_1 to signify the instantaneous (employment) growth rate of the industry over a period if growth is positive, otherwise zero;[11] x_2 is the absolute value of the instantaneous growth rate of the industry over a period if growth is strictly negative, otherwise zero. These two growth variables are entered as explanatory variables in the regression equation for entry in the form $b_1x_1^{a_1} + b_2x_2^{a_2}$, where the a's and b's are parameters to be estimated. Our theory implies that entry is an increasing function of positive growth and suggests that this function is convex. Hence we expect $b_1 > 0$ and the exponent $a_1 > 1$. We also expect $b_2 < 0$, since a high negative growth rate of the industry (i.e., industrial contraction) indicates low profitability and should discourage entry. We have no hypothesis about the exponent a_2 beyond $a_2 > 0$.

---

[8] A formal proof that g_e is an increasing convex function of g_x is straightforward, given the assumed properties of $j_o(g_o)$ ($j_o' > 0, j_o'' > 0$) and of $j_e(g_e)$ ($j_e' > 0, j_e'' < 0$). By the implicit function theorem, the inverse functions $g_o = h_o(P_K)$ and $g_e = h_e(P_K)$ exist and have derivatives satisfying the conditions $h_o' > 0$; $h_o'' < 0$, $h_e' > 0$; and $h_e'' > 0$. Since $g_x = g_o + g_e$ is an identity (given the assumption of no firm exit), we obtain the identity $g_x = h_x(P_K) = h_o(P_K) + h_e(P_K)$, with $h_x' > 0$. Using the chain rule, we obtain $dg_e/dg_x = h_e'/h_x' > 0$ and $d^2g_e/dg_x^2 = (h_o'h_e'' - h_e'h_o'')/(h_x')^3 > 0$. Thus g_e is an increasing convex function of g_x.

[9] The extension allowing for exit assumes a density function of capital $\psi(P_K)$ that will be withdrawn by exiting firms, depending on their opportunity cost of remaining in the industry. Then the (negative) component of the industry growth rate due to exit is $g_d = \int_{P_K}^{\infty}\psi(x)dx$. If P_K is always greater than location of the mode of this density function, it follows that the function $g_d = h_d(P_K)$ is a decreasing convex function, i.e., $h_d' < 0$ and $h_d'' > 0$. The assumptions on $h_o(P_K)$ and $h_e(P_K)$ are in n. 8. We have the identity $g_x = g_o + g_e - g_d = h_x(P_K)$, with $h_x' > 0$. By the chain rule, $dg_e/dg_x = h_e'/(h_o' + h_e' - h_d') > 0$ and $d^2g_e/dg_x^2 = [(h_o' - h_d')h_e'' - h_e'(h_o'' - h_d'')]/(h_x')^3 > 0$. Hence g_e is still an increasing convex function of g_x under our assumptions even when exit is taken into account. By a similar argument, it can be shown that g_d is a decreasing function of g_x, but its convexity or concavity is undetermined without further assumptions.

[10] Another limitation of the model is its partial equilibrium structure. What is the nature of the pool of potential entrants and their resources? If the shadow price of capital is high enough to produce rapid growth over a number of periods, should one expect a change in the relative shares of internal expansion and entry? It is difficult to pursue these questions without additional empirical evidence on appropriate restrictions that should be imposed.

In our theoretical framework, entry rates may also be influenced by interindustry differences in the characteristics of the pool of potential entrants and/or in the implicit costs of entry (relative to the costs of expansion by old firms). We consider briefly several categories of such interindustry differences, how they would affect entry, and the variables used as empirical counterparts.

1. The minimum size of plant required for efficient entry has been widely discussed. There are several reasons why major economies of scale are likely to retard entry. First, the number of potential entrepreneurs possessing the capital requirements presumably declines as the capital requirements increase, and the supply curve of capital to them is not perfectly elastic. This consideration is especially relevant for new firm entry. Second, it is plausible that larger plants incorporate relatively more specialized capital resources, and therefore the pool of preexisting resources that can be elastic
such industries is smaller.[12]

The minimum efficient size of plant is also a
the well-known Modigliani–Sylos Labini limit
ample, in Modigliani (1958). In this model, a
above the competitive level can arise without a
mies of scale are so great that an efficient-sized
to size of the industry. Studies of entry and industry
include an explanatory variable indicating the size of an efficient plant or firm relative to the industry, based on this line of argument.

In our empirical work, we include variables x_3 and x_4 as measures of the absolute and relative size of efficient new plants in an industry.

---

[11] Let m_s and m_t denote industry employment at times s and t. The variable x_1 [log (m_t/m_s)]/$(t - s)$ if positive, otherwise zero.

[12] There is another explanation why x_4 may be negatively correlated with entry. Hause (1962) argued that the expected life of small but established firms is shorter than the expected life of large firms. Hence entry may be high in industries where average firm size is small because of turnover of small firms. Since average firm size is probably positively correlated with the minimum optimal entry size, the conclusion follows.

The variable x_3 is the logarithm of the average size of new establishments built by new, diversifying, and old firms already established in the industry. Size is measured by employment at the end of the subperiod in which the establishment was built. However, an average is taken of the three subperiods of data, to reduce the sampling variation. This variable is intended as a crude proxy for the optimum size of new establishments. In previous studies, a number of unconvincing and extremely ad hoc proxies have been used, for example, average size of all firms or plants in the industry or size of plant such that 50 percent of industry output comes from larger plants. Since entrepreneurs building new establishments have nontrivial incentives to build economically viable plants, our proxy has some appeal.[13] We expect this variable to be negatively correlated with entry. We use x_4 to signify the size of an efficient new plant relative to 1968 industry employment. It is measured by the logarithm of the ratio of an efficient-sized new plant (as defined in the preceding paragraph) to 1968 industry employment. Large values of this variable indicate that an industry is not large enough to contain many optimum-size new plants. If Modigliani's argument is relevant for our data, one expects a negative relationship between entry and x_4.

2. Our model of the entry-growth relationship is essentially competitive and abstracts from strategic firm behavior when significant long-term monopoly power exists in the industry. Since monopoly power implies the inability to restrict industry output, the existence of long-term monopoly power will presumably be associated with less entry almost by definition. An unusual variable is available as an indicator of monopoly for this study. The variable x_5 is a dummy variable with a value of one if the industry has a significant registered

---

[13] Let A_n, A_d, and A_o be the average size of new establishments built by new firms, diversifying firms, and old firms already established in the industry. For industries where the comparisons can be made, $A_n < A_d$ 23 out of 24 times, and $A_n < A_o$ 30 out of 31 times. But $A_d < A_o$ only 13 out of 22 times. These findings provide strong evidence that new firm establishments are relatively small. But the comparison of A_d and A_o provides much less evidence of a capital constraint for diversifying firms relative to old firms already in the industry for the construction of new plants. The table of means and standard deviations of the log ratios of these average sizes across industries at the end of this note provides further evidence. A rather ingenious way of measuring minimum optimum size, based on multiple establishment data of old firms in the industry, was recently proposed by Lyons (1980). This suggestion was discovered too late for consideration in this study

	Mean	Standard Deviation
log (A_d/A_n)	1.20	.71
log (A_o/A_n)	1.48	.88
log (A_o/A_d)	.29	1.04

cartel agreement, otherwise zero. This classification of industries is reported in Carling (1968). With one exception these cartel agreements were made well before the beginning of the sample period in 1954. Consequently we interpret this variable as an indicator of a durable noncompetitive element in the market and expect it to be negatively associated with entry. The appropriate causal interpretation of this variable (or for any variable indicating monopoly power) is unclear. The presence of a durable cartel agreement might indicate deliberate cartel policies to retard entry. But it could also be a proxy indicating the presence of other unmeasured obstacles to entry, and these obstacles might be a necessary condition for the long-term viability of the cartel.

It seems plausible that entry will play an especially important role in industries experiencing rapid technological change or selling new products. Some calculations not reported in this study were carried out with very crude proxy variables representing these factors. These calculations provide some statistical support for this conjecture. Some of these results are published in Du Rietz (1980). We think this topic is an interesting one that deserves further theoretical and empirical study.

IV. Econometric Specification and Estimation of New Firm Entry

The empirical analysis is restricted to a single-equation reduced form We discuss the specification of this equation, the data used in the analysis, and the results from estimating homosced
scedastic versions of the equation.

Entry is measured empirically by the ratio
firms at the end of a period to total indus
beginning of the period, unless otherwise
equation regression model is

$$\mathbf{y} = b_0 + b_1 \mathbf{x}_1^{a_1} + b_2 \mathbf{x}_2^{a_2} + \sum_{i=3}^{5} b_i \mathbf{x}_i$$

In this equation, **y** is the vector of 117 observations from 39 manufacturing industries (most of the plastics, basic metals, and engineering

---

[11] Input (employment) measures of entry and of the industry growth rate are used since they reduce the statistical noise from absolute and relative price changes that would occur if nominal output measures, such as sales or value added, were used. Most empirical studies of entry have measured entry by the ratio of new to old firms except for Hause (1962) and Du Rietz (1975, 1980) We prefer our measure as an indicator of the direct quantitative importance of entry, since new firms are on average very much smaller than existing firms.

industries in Sweden) for the three periods 1954–58, 1959–63, and 1964–68. The independent variables have been defined in the preceding section. They include positive industry growth rates of employment, x_1; absolute value of negative industry growth rates of employment, x_2; log of average employment of newly constructed plants (indicator of efficient size of new plants), x_3; log of ratio of x_3 to 1968 industry employment (indicator of relative size of efficient new plant), x_4; and a dummy variable for industries with significant cartel agreements (indicator of monopoly power), x_5.

The disturbance term, **u**, in equation (5) is assumed independently and normally distributed. Two alternative specifications are considered. In the first, the variance of **u** is assumed constant. In the second, we adopt the model of multiplicative heteroscedasticity proposed by Harvey (1976). This model assumes that the variance of each observation is

$$\sigma_{ii} = \exp(\mathbf{z}_i'\mathbf{c}). \qquad (6)$$

In (6), $\mathbf{z}_i$ is a vector of observed variables related to the ith observation and **c** is a corresponding vector of parameters to be estimated. In cross-sectional industry studies, outlying observations that greatly affect the least-squares fitting are often observed. The residuals from these outliers are often so large that they make the distributional assumptions of independently and identically distributed normal errors untenable. Harvey's model is an appealing specification. It assumes that the intrinsic variability of the dependent variable is a function of observable characteristics of each observation. Furthermore, those characteristics associated with high variability may be of substantive economic interest. In this study, we consider for **z** a two-variable subset of the **x** variables that explain the mean part of the model. These variables include positive industry growth, x_1, and new plant optimum size, x_3. Industries experiencing high growth rates are plausibly associated with stochastically noisier environments than those with low growth. Our entry measure has a lower bound of zero, and we expect that the variance of entry will be positively correlated with its level. The optimum size of new plants should be negatively correlated with new firm entry, and so small values for x_3 are likely to be associated with both high entry and high variance. The specification of the variance in Harvey's model has some pitfalls for the unwary and is discussed further in Appendix B.

The results of the first two regression equations are reported in table 2. They indicate that OLS and the Harvey heteroscedastic model yield qualitatively similar results, since variable parameters have iden-

TABLE 2

OLS and Harvey Regression Parameter Estimates for New Firm Entry

		Independent Variable Coefficients*								
		x_1		x_2						
Number	x_0	Linear	Exp	Linear	Exp	x_3	x_4	x_5	R^2	Log Likelihood
1. OLS mean	7.83	.055	2.18	−1.07	440	−1.81	−0165	−683	46	−176
	(2.6)	(060)	(46)	(76)	(36)	(43)	(31)	(67)		
2. Harvey mean	6.33	00300	3.13	−69	507	−1.32	−0228	−702	33	−156
	(1.5)	(011)	(1.7)	(27)		(23)	(19)	(40)		
Var	6.46	261				−1.68			..	. .
	(67)	(050)				(20)				
3. Harvey mean	6.04	201		−216		−1.33	−0221	−731		−157
	(1.5)	(11)		(10)		(26)	(19)	(41)		
Var	6.46	261				−1.68		.	..	...
	(1.2)	073				(36)				
4. OLS mean	12.4			1.70	383	−1.47	−0581	−3.83	24	−322
	(9.3)			(3.1)	(87)	(1.6)	(1.1)	(2.4)		
5. Harvey mean	13.7			−2.19	3804	−1.36	0151	−4.76	.	−302
	(1.9)			(1.1)		(1.0)	(73)	(1.5)		
Var	6.0				.	−821	.		..	
	(1.6)					(20)				

Note.—The R^2 for the Harvey [illegible] least squares regressions with weights obtained from the variance estimate from the model. Numbers in parentheses are asymptotic standard errors.

* The independent variable [illegible] growth rate of industry, x_3 = negative growth rate of industry (absolute value), x_4 = proxy for optimum size of new plants: log of average size of new [illegible] new plant to industry employment, and x_5 = dummy variable for significant registered cartel agreement. Linear and exponential parameter [illegible] in regressions 1–3 is new firm entry rate measured by employment; in 4 and 5 it is measured by numbers of firms.

† The exponent was [illegible] of difficulty in obtaining convergence.

tical signs in both regressions.[15] All regression coefficients have the theoretically expected signs. A positive industry growth rate (x_1) is associated with higher entry rates. A negative industry growth rate (x_2), an increase in the minimum efficient size of new establishments (x_3), and the existence of a significant registered cartel agreement (x_5) are all associated with a lower rate of entry. Minimum efficient size of new plant relative to industry employment, x_1, has the predicted negative sign, but it is trivial in size and statistical significance. Hence these data provide no support for the Modigliani–Sylos Labini model of relative plant size barriers to entry.

Both regressions 1 and 2 qualitatively support the hypothesis that entry is an increasing, convex function of the industry growth, since the point estimates of $b_1 > 0$ and the point estimates of the exponential parameter $a_1 > 1$. The asymptotic t-value for the null hypothesis that $a_1 = 1$ is 2.3 in the OLS regression 1, while the corresponding t-value in the heteroscedastic regression 2 is more modest, 1.22, just short of the 10 percent significance level for a one-tailed test. Three comments should be made on these statistical results. First, the data strongly reject the OLS hypothesis of homoscedasticity, and so statistical inference should be based on regression 2, not regression 1. This issue is discussed below. Second, the lower t-value for the null hypothesis $a_1 = 1$ in regression 2 is probably a consequence of using industry growth (x_1) as an explanatory variable of the heteroscedasticity in the Harvey model. Third, it is difficult to obtain a sharp test of the convexity of the entry function because of the extremely high sampling correlation (−.990) between the linear growth parameter b_1 and the exponential growth parameter a_1. This correlation is a consequence of trying to allocate the effect of x_1 on entry through two distinct parameters and is analogous to severe multicollinearity. Given this statistical problem, we consider that regression 2 provides reasonable if modest support for the convexity hypothesis. The small t-value for b_1 also stems from this correlation and does not imply that x_1 is a statistically insignificant explanatory variable.[16]

[15] This similarity of parameter signs for the two regressions is not surprising. The mean part of Harvey's regression model is equivalent to a weighted least-squares regression with weights $(1/\sigma_u)^{1/2}$, obtained by using the variance parameter estimates from Harvey's model. Unless the weights are very different, the coefficients obtained by OLS and weighted least squares usually have the same sign.

[16] Regression 3 reports the results if the growth rates are entered linearly. Since no exponential parameter competes with the linear parameter b_1 for transmitting the effect of x_1 on the entry rate, the parameter b_1 is statistically significant in this regression. Allowing for convexity improves the overall regression fit very little. The log of the likelihood ratio of regressions 3 and 2 times 2 is 1.64, which is asymptotically distributed χ^2 with 1 *df*. Hence the hypothesis of a linear relationship of growth on entry is rejected at the 22 percent level. However, the relevant hypothesis is whether $a_1 > 1$, not the two-tailed test whether $a_1 \neq 1$. Hence, the relevant significance level is

Although the signs of estimated parameters are identical in regressions 1 and 2, there are substantial discrepancies both in the magnitudes of some parameters and in their calculated t-values. The parameter estimates often differ by a factor of 1.5. The cartel dummy x_7 has a significant negative relationship with entry by conventional standards in the Harvey model, but not in OlS, and the OLS t-value for the exponential coefficient, a_1, exaggerates the statistical importance of this parameter.

The data overwhelmingly reject the assumption of homoscedasticity, the hypothesis underlying the OLS calculation of regression 1. The log of the likelihood ratio of the restricted regression 1 to the unrestricted regression 2 times -2 is 40. This statistic has a χ^2 distribution with 2 degrees of freedom under the null hypothesis, and hence the null hypothesis is decisively rejected. This implies that the estimates and statistical inferences based on OLS regression 1 are completely unacceptable.

The relationship between regressions 1 and 2 can best be understood by recalling that the mean part of the Harvey model is equivalent to a weighted regression. The weights are equal to the reciprocal of the estimated standard deviation for each observation, using the parameters obtained for the variance part of the model. The coefficient of variation of the weights is .62, and the ratio of the largest to smallest weight is about 24. These summary statistics on the weights indicate that the appropriate heteroscedastic weights depart substantially from the uniform weighting of the OLS regression and explain why these regressions differ. Closer insp [illegible] reveals that the parameters of each variable th [illegible] mean and variance part of the regression hav [illegible] implies that the variance of entry is positively co [illegible] of entry. The correlation of the weights with the [illegible] entry rate is $-.68$, while the correlation of t [illegible] entry is $-.42$. These large correlations confirm t [illegible] of the Harvey model in reducing the importance of [illegible] rate observations in the regression.

These results indicate that inference based on OLS with cross-sectional industry data is hazardous and can be extremely misleading in the absence of a proper test of homoscedasticity. Failure to correct for severe heteroscedasticity can thoroughly vitiate the statistical results of an otherwise well-executed empirical study.

We therefore restrict further quantitative discussion of the parameter estimates to the Harvey regression 2. The size of the effect of

---

about 11 percent. This result is very close to our direct test of $a_1 = 1$ by its asymptotic t-value in regression 2.

optimum new plant size on entry implied by the parameter of x_3 is substantial. Since the log of this variable was used in the regression, the coefficient indicates that a doubling of the size leads to an expected reduction in the annual entry rate of about -1.3 percentage points. The t-value for this coefficient is rather high, about 4. The presence of a strong cartel agreement (x_5) is associated with a reduction of the entry rate by 0.7 percent, an empirically modest amount, but statistically significant (5 percent level). The proper causal interpretation of this correlation is unclear. Do cartels make entry more difficult, or does the lack of entry facilitate cartels?

What happens if entry is measured by the ratio of new to initial numbers of firms instead of the ratio of new firm employment to initial industry employment? Regressions 4 and 5 are the counterparts of regressions 1 and 2 in table 2, using this alternative measure of entry. This entry measure behaves somewhat erratically and has a larger coefficient of variation than the employee entry measure. Even so, there is qualitative similarity of these new regressions with 1 and 2. Homoscedasticity is again emphatically rejected, since -2 times the log likelihood ratio has a value of 40, and the statistic is asymptotically χ^2 with 2 degrees of freedom under the hypothesis of homoscedasticity. We therefore limit our discussion to a comparison of the heteroscedastic regressions 5 and 2.

All parameters in these two regressions have the same sign, except the trivially small and statistically insignificant coefficient of x_1. The magnitude of the convexity (exponential) parameter a_1 is about 25 percent smaller but gives slightly more statistical support to the hypothesis that $a_1 > 1$. The statistical support for the significance of x_3 is much weaker, but the statistical significance of the negative association of cartel agreements and entry is much stronger. These differences are partially accounted for by the smaller effect that x_3 has in determining the heteroscedasticity in 5, which in turn alters the observation weights in determining the mean part of the regression. Although both definitions of entry lead to similar patterns of results, we have greater confidence in the former, based on employment of new firms.

In summary, these results certainly suggest responsiveness of new firm entrants to changes in costs and returns, which provides some evidence supporting belief in the importance of potential competition. There is strong statistical evidence that new firm entry is positively related to industry growth. There is more modest statistical support for the conclusion that entry is a convex function of growth in this sample.

The average size of new plants built in an industry seems to be a sensible indicator of the size of resource commitment required for entry. This variable has a strong negative association with new firm entry. There is also significant support for concluding that monopoly power (as indicated by cartel agreements) has a negative association with entry. The causal basis of this relationship is undetermined.

Finally, our statistical work indicates the great importance of examining cross-section industry data for heteroscedasticity and taking it into account before attempting statistical inference (or estimation). Otherwise, one is likely to draw ill-founded conclusions. We have far more confidence in the results from Harvey's heteroscedastic model than from the OLS calculations. However, our experience suggests that Harvey's model is a delicate one, requiring careful specification. The development of more robust models for dealing with heteroscedasticity will be important for more reliable empirical conclusions.

Appendix A

Dynamic Optimization by a Competitive Firm[17]

The firm's objective is maximization of the present value of its cash flow:

$$\int_t^{\infty} [PX(K, L, I) - WL - GI]e^{-rs}ds. \qquad (A1)$$

The production function incorporates costs of adjustment of the capital stock as given by (1) in the main text, $X(K, L, I) = F(L, K) - KC(I/K)$. The firm has rational expectations about the product price P and factor prices W and G. These prices may vary with time. The constraints are the firm's capital stock at t ($K[t] = K_t$), the gross and net investment identity ($\dot{K} = I - mK$, where m is the exogenous depreciation rate), and the boundary conditions (K, $L > 0$). The firm chooses $L(s)$ and $I(s)$ ($t < s < \infty$) to maximize (A1). We adopt the Hamiltonian approach for solving this problem.

The current value Hamiltonian is

$$H(K, L) = P\left[F(K, L) - KC\left(\frac{I}{K}\right)\right] - WI$$

Necessary conditions for a maximum include

$$\frac{\partial H}{\partial I} = 0, \quad \text{so } (-PC' - G + \lambda$$

$$\frac{\partial H}{\partial L} = 0, \quad \text{so } (PF_L - W) = 0.$$

$$\dot{\lambda} = \frac{-\partial H}{\partial K} + r\lambda, \quad \text{so } \dot{\lambda} = (r + m)\lambda - P\left(F_K + \frac{IC'}{K} - C\right). \qquad \text{(iii)}$$

$$\lim_{T \to \infty} \lambda(T)e^{-rT} = 0 \text{ (transversality condition with } \infty \text{ horizon)}. \qquad \text{(iv)}$$

$$\dot{K} = I - mK \qquad \text{(v)}$$

---

[17] Ken Hendricks kindly supplied the details of the following analysis including the distinction between the Hamiltonian and calculus of variations formulations of the problem.

The costate equation is

$$\begin{aligned}\dot{\lambda} - (r + m)\lambda = f(t) &= -P\left(F_K + \frac{IC'}{K} - C\right)\\ &= -PX_K.\end{aligned} \tag{A3}$$

The formal solution to the costate equation is

$$\frac{d[\lambda e^{-(r+m)t}]}{dt} = f(t)e^{-(r+m)t}. \tag{A4}$$

Integrating (A4) from t to T, letting T approach ∞, and using the transversality condition (iv) yields

$$\begin{aligned}\lambda(t) &= \int_t^{\infty} PX_K e^{-(r+m)(s-t)}ds\\ &= PC' + G\end{aligned} \tag{A5}$$

(from the necessary marginal condition [i]). The last pair of equations and equation (ii) are the marginal conditions for the optimum cited in the main text.

If one adopts the calculus of variations approach and uses $K(s)$ as a control variable instead of $I(s)$, one obtains the marginal condition

$$\int_t^{\infty}(PX_K - Gm)e^{-r(s-t)}ds = PC' + G. \tag{A6}$$

Equations (A5) and (A6) look slightly different but are both correct and make economic sense when one considers the difference between gross and net investment. But the interpretation of (A5) is perhaps more immediately obvious.

Appendix B

A Note on Correcting Heteroscedasticity with Harvey's Model[18]

Harvey's heteroscedastic model (1976) allows the variance of each observation to depend on observation-specific variables, by assuming the ith observation is distributed:

$$y_i \sim N[\mathbf{x}_i'\mathbf{b}, \exp(\mathbf{z}_i'\mathbf{c})] \tag{B1}$$

In this model, $\mathbf{x}_i$ and $\mathbf{z}_i$ are $m \times 1$ and $p \times 1$ vectors of observed covariates determining the mean and variance of y_i, respectively. They may contain identical, distinct, or overlapping sets of variables. ML estimates of the parameter vectors **b** and **c** may be estimated iteratively from the equations

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{B2}$$

and

$$(\mathbf{1}_n - \mathbf{r})'\mathbf{Z} = \mathbf{0}_p. \tag{B3}$$

---

[18] See Godfrey (1978), Breusch and Pagan (1979), Carroll and Ruppert (1982), and Kao (1983) for further discussion of the model.

V in (B2) is the $n \times n$ diagonal matrix with ith entry $\exp(\mathbf{z}_i'\mathbf{c})$, the parametric estimate of the variance of y_i. Hence (B2) is the familiar WLS estimator, with weights determined by the estimated variances of the individual y_i. In (B3), $\mathbf{1}_n$ is the $n \times 1$ vector of ones, $\mathbf{0}_p$ the $p \times 1$ vector of zeroes, and **Z** the $n \times p$ observation matrix of the z's; **r** is an $n \times 1$ vector with ith entry $[s_i^2/\exp(\mathbf{z}_i'\mathbf{c})]$, with s_i the residual of y_i. Consistent initial estimates of **b** are obtained from OLS regression of y on x. Consistent initial estimates of **c** are obtained from OLS regression of log (s^2) on z, where s^2 is the squared residual from the preceding calculation.

The covariance matrices for the estimated **b** and **c** parameters are

$$\operatorname{var}(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{B4}$$

and

$$\operatorname{var}(\hat{\mathbf{c}}) = 2(\mathbf{Z}'\mathbf{Z})^{-1}. \tag{B5}$$

Equation (B4) is the usual WLS covariance matrix for the **b** parameters. Equation (B5) shows that the asymptotic covariance matrix of the estimated **c** parameters depends only on the observed z's, and not on any parameters of the model. This last relationship reveals a defect of Harvey's model. It assumes that the specification of the individual variances in (B2) is exactly correct. In OLS, we assume that the residual subsumes misspecification, and the size of the variance is another parameter to be estimated. Harvey's model has no scale parameter in the variance part of the model (B2), to allow for similar misspecification. This defect may be important when inference on specific parameters of **c** is an issue. This problem seems to have been ignored in the published literature on Harvey's model

There is a pitfall in using Harvey's model analogous to a problem that arises when severely heteroscedastic data are uncorrected in an OLS regression analysis. With small or moderate size samples, OLS estimation and inference can be unduly influenced by a few extreme observations with very large residuals. Similarly, in Harvey's model, observations yieldin
uals can greatly influence the estimates of the **c** pa
vations in turn will tend to be given much weight
Thus, in one specification not reported above, a va
mean and variance parts of the model yielded pa
surdly large t-values (-7.3 and -16). Given the p
liers" in the weights from Harvey model estimat
reporting of the coefficient of variation of the wei
imum to minimum weight.

Despite this potential delicacy of the Harvey model, w
major improvement over OLS calculations with no test for heteroscedasticity. The LR test decisively rejects homoscedasticity in this study, and we find large differences in the OLS and Harvey estimates of the **b** (mean) parameters and their nominal statistical significance. Heteroscedasticity is a serious problem in many cross-sectional industry studies. Failure to recognize the problem or to take it appropriately into account leaves all statistical estimates and hypothesis testing at great risk

References

Bain, Joe S. "Pricing in Monopoly and Oligopoly." *A E R* 39 (March 1949): 448–64.

Baumol, William J., Panzar, John C.; and Willig, Robert D. *Contestable Markets and the Theory of Industry Structure.* New York: Harcourt Brace Jovanovich, 1982.

Breusch, T. S., and Pagan, Adrian R. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47 (September 1979): 1287–94.

Brock, William A., and Scheinkman, José A. "On the Long-Run Behaviour of a Competitive Firm." In *Equilibrium and Disequilibrium in Economic Theory,* edited by Gerhard Schwodiauer. Dordrecht: Reidel, 1978.

Carling, Alf. *Industrins Struktur och Konkurrensforhållanden: Koncentrationsutredningen III.* Stockholm: Statens Offentliga Utredningar, 1968.

Carroll, R. J., and Ruppert, David. "A Comparison between Maximum Likelihood and Generalized Least Squares in a Heteroscedastic Linear Model." *J. American Statis. Assoc.* 77 (December 1982): 878–82.

Clark, John Bates. "The Limits of Competition." *Polit. Sci. Q.* 2, no. 1 (1887): 45–61.

Duetsch, Larry L. "Structure, Performance, and the Net Rate of Entry into Manufacturing Industries." *Southern Econ. J.* 41 (January 1975): 450–56.

Du Rietz, Gunnar. *Etablering, Nedläggning och Industriell Tillvaxt i Sverige, 1954–1970.* Stockholm: Industriens Utredningsinstitut, 1975.

———. *Foretagsetableringarna i Sverige under Efterkrigstiden.* Stockholm: Industriens Utredningsinstitut, 1980.

Godfrey, Leslie G. "Testing for Multiplicative Heteroskedasticity." *J. Econometrics* 8 (October 1978): 227–36.

Gorecki, Paul K. "The Determinants of Entry by New and Diversifying Enterprises in the U.K. Manufacturing Sector, 1958–1963: Some Tentative Results." *Appl. Econ.* 7 (June 1975): 139–47.

Gould, John P. "Adjustment Costs in the Theory of Investment of the Firm." *Rev. Econ. Studies* 35 (January 1968): 47–55.

Harvey, A. C. "Estimating Regression Models with Multiplicative Heteroscedasticity." *Econometrica* 44 (May 1976): 461–65.

Hause, John C. "New Firm Entry in Manufacturing Industries." Ph.D. dissertation, Univ. Chicago, 1962.

Holt, Charles C.; Modigliani, Franco; Muth, John F.; and Simon, Herbert A. *Planning Production, Inventories, and Work Force.* Englewood Cliffs, N.J.: Prentice-Hall, 1960.

Kao, Chih-Hwa. "Essays on the Estimation of Regression Models with Multiplicative Heteroscedasticity." Ph.D. dissertation, SUNY at Stony Brook, 1983.

Lucas, Robert E., Jr. "Adjustment Costs and the Theory of Supply." *J.P.E.* 75, no. 4, pt. 1 (August 1967): 321–34.

Lyons, Bruce R. "A New Measure of Minimum Efficient Plant Size in U.K. Manufacturing Industry." *Economica* 47 (February 1980): 19–34.

McGuckin, Robert. "Entry, Concentration Change, and Stability of Market Shares." *Southern Econ. J.* 38 (January 1972): 363–70.

Mansfield, Edwin. "Entry, Gibrat's Law, Innovation, and the Growth of Firms." *A.E.R.* 52 (December 1962): 1023–51.

Modigliani, Franco. "New Developments on the Oligopoly Front." *J.P.E.* 66 (June 1958): 215–32.

Mussa, Michael L. "External and Internal Adjustment Costs and the Theory of Aggregate and Firm Investment." *Economica* 44 (May 1977): 163–78.

Nickell, S. J. *The Investment Decisions of Firms.* Welwyn: James Nisbet; Cambridge: Cambridge Univ. Press, 1978.

Orr, Dale. "The Determinants of Entry: A Study of the Canadian Manufacturing Industries." *Rev. Econ. and Statis.* 56 (February 1974): 58–66.

Rothschild, Michael. "On the Cost of Adjustment." *Q.J.E.* 85 (November 1971): 605–22.

Smith, Adam. *An Inquiry into the Nature and Causes of the Wealth of Nations* 1776. New York: Modern Library, 1937.

Soderström, Hans T. "Team Production and the Dynamics of the Firm." *Scandinavian J. Econ.* 79, no. 3 (1977): 271–88.

Stigler, George J. *The Organization of Industry* Homewood, Ill.: Irwin, 1968.

Treadway, Arthur B. "On Rational Entrepreneurial Behaviour and the Demand for Investment." *Rev. Econ. Studies* 36 (April 1969): 227–39.

———. "Adjustment Costs and Variable Inputs in the Theory of the Competitive Firm." *J. Econ. Theory* 2 (December 1970): 329–47.

Wedervang, Frøystein. *Development of a Population of Industrial Firms: The Structure of Manufacturing Industries in Norway, 1930–1948* Oslo. Universitetsforlaget, 1965

Job Matching with Finite Horizon and Risk Aversion

Milton Harris

Northwestern University

Yoram Weiss

Tel-Aviv University

The consequences of introducing risk-averse workers with a known, finite retirement age into a simple job-matching framework are explored. The job-matching model includes two jobs. In the "primary" job, workers and firms learn about the worker's productivity by observing his output over time. In the "secondary" job, productivity of any worker is known with certainty at the outset. Workers start off in the primary job but, based on their observed performance, may choose to switch to the secondary job. The model is developed in two stages. First, the finite retirement age is introduced while the usual assumption of risk-neutral workers is retained. In this model it is shown that all workers who achieve at least a certain cumulative output record, denoted r^*, by a certain age, denoted T, will thereafter remain in the primary job until retirement. Those who do not achieve r^* by age T switch to the secondary job at or before age T and remain there. Second, risk aversion is introduced along with finite retirement age. Here the turnover aspects of the equilibrium are similar to the risk-neutral case, but firms provide workers with wage insurance. This takes the form of wages that are downward rigid. As a result, it may be optimal ex ante for workers to delegate the job choice decision to the employer, and, ex post, not all separations will be voluntary for the worker.

We are indebted to Truman Bewley, J. Michael Harrison, James Heckman, Bengt Holmstrom, Paul Milgrom, Rick Miller, Roger Myerson, Paul Pfleiderer, two anonymous referees, and participants of the Uncertainty Workshop, Graduate School of Business, University of Chicago, for very helpful input. Support from the National Science Foundation is gratefully acknowledged.

[*Journal of Political Economy*, 1984, vol. 92, no. 4]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9204-0002$01.50

I. Introduction

The job-matching literature has generally focused on models in which workers are infinitely lived and risk neutral (e.g., Mortensen 1978; Jovanovic 1979; Viscusi 1979). In this paper we explore the consequences of reversing these two assumptions. First we consider a simple job-matching framework with risk-neutral workers who have a known, finite retirement age. Second we analyze the same framework but assume workers are risk averse. The results are considerably different from those of conventional job-matching models, as will be discussed after a brief description of the models.

The basic job-matching framework is as follows. Workers may select one of two "jobs," which we will refer to as the "primary" job (since workers will be assumed to try this one first) and the "secondary" job. Workers are heterogeneous with respect to their productivity in the primary job. Both workers and firms are assumed to be unsure of any given worker's productivity in the primary job, but the productivity of all workers in the secondary job is common knowledge. Output in the primary job is random, but its distribution depends on the worker's true productivity in that job. In particular, output is assumed to follow a Poisson arrival process in which the average arrival rate is equal to the worker's true productivity. Both workers and firms update their beliefs about the worker's productivity by observing his output in the primary job over time. Wages are determined by a zero-expected-profit condition on the risk-neutral firms. In the secondary job, this implies that wages equal (known) productivity. Given the way productivity estimates, and hence wages, will evolve stochastically in the primary job, workers
each age, which job to select (moving betwee
possible but costly).[1] This basic job-matching s
two models.

Initially (in Sec. II) we retain the standar
neutral workers and explore in detail the effe
workers have the same, exogenous, determin
age. Under risk neutrality, each worker currently in
chooses whether to switch jobs at each age so as to maximize his expected wealth. It will be shown that once a worker selects the secondary job, he never returns to the primary job. Workers with sufficiently favorable records will choose to remain in the primary job for the rest of their career, irrespective of their future performance. Others will be forced by circumstances (e.g., their poor performance

---

[1] The problem of a worker in choosing a job is similar to the "Two-armed Bandit" problem first discussed in the economics literature by Rothschild (1974). This fact will be used below.

record) to switch to the secondary job. Because of risk neutrality of workers, the firm does not provide insurance and is not bound to retain any worker whose expected contribution to the firm's profits is negative. Wages at each instant are equal to the perceived marginal productivity of the worker, where perceptions are determined by the observed cumulative output record, r.

The main result is that there exists an increasing sequence of "optimal stopping times," $t(r)$, $r = 0, 1, 2, \ldots$, where $t(r)$ is the longest a worker with record r would remain in the primary job. Moreover, we show that there is some finite r^* such that $T = t(r^* - 1) < A$ and $t(r) = A$ for all $r \geq r^*$, where A is retirement age. That is, workers in the primary job with record $r^* - 1$ at age T will switch to the secondary job while those with records of r^* or better will never switch. Intuitively, once a worker achieves a record of r^*, even if his record thereafter is poor, his perceived productivity does not decline as fast as his moving costs increase due to the finite lifetime. In the primary job, a worker's wages may fluctuate both up and down as his productivity estimate is revised, and, of course, a worker may choose to switch to the secondary job before T.

In contrast to Jovanovic (1979), who deals with the case in which information is firm (or match) specific, our model analyzes the case in which information is worker specific. Therefore, as in Viscusi (1979), a worker whose productivity in the primary sector is revealed to be low will turn to an alternative sector (i.e., to a different stochastic process) rather than sample again from the same process. Our model is therefore applicable to the empirical analysis of interindustry movements in the labor market. It suggests that if a worker can choose between an occupation in which experimentation is important and an occupation in which skills are quickly revealed, he will initially choose the risky occupation (or never try it). Mobility will be from the risky occupation into the less risky one. The age pattern of mobility flows is similar to the *intra*industry pattern discussed by Jovanovic. As the worker becomes older, he is less likely to move. The reasons are, however, somewhat different in the two models. In Jovanovic, as the match continues, the worker becomes more certain of its quality. Thus given that the worker is still with the firm, the probability of extreme (low) values declines, and quitting is less likely. In our model, the worker has a finite life, which implies that mobility also declines with age given tenure with the firm. This prediction is supported by the findings of Jovanovic and Mincer (1978) from the Michigan Income Dynamics Survey.

The turnover patterns produced by our model are similar to those observed in many industries, especially academia (indeed we will refer to r^* as the tenure standard and T as the up-or-out age in the sequel).

Thus this simple job-matching model with finite retirement age can rationalize some of the turnover aspects of the academic labor market. It should be noted, however, that in this model "tenure" provides no wage insurance, and all separations are completely voluntary. In this sense, the employment contract does not bind the firm or the worker. In observed contracts, however, firms are obliged to retain some workers at wages such that their net contribution is negative, and some workers are obliged to leave when, ex post, they would prefer not to. The lack of insurance is clearly related to the assumption of risk-neutral workers, which also accounts for the fact that separations are voluntary. This leads us to explore the consequences of introducing worker risk aversion.[2]

In the second part of the paper, workers are risk averse and have no access to the capital market.[3] The model is very similar to those of Freeman (1977), Weiss (1981), and Harris and Holmstrom (1982). In those models, however, the worker's output process is independent of his job; that is, all jobs are identical and there is no endogenous turnover. In our model, employment and salary strategies are derived from the maximization of the worker's expected utility given the expected profits of the firm (we retain the assumption that firms are risk neutral). The behavior of wages in this case reflects the desire of workers to smooth consumption over time and across different realizations of the output process. Wages do not always follow productivity, and, in particular, a worker who is retained by the firm will *not* suffer a reduction in wages just because his output record is not favorable. (Freeman [1977], Weiss [1981], and Harris
[1982] also obtain this result, but not in a job
result of this wage rigidity, a worker may
even though his expected productivity (as
below his productivity elsewhere. To
inefficiencies, the worker agrees, ex ante, to

---

[2] There are, of course, many other aspects of observed
not captured in our simple model. For example, it is often said
freedom of expression and that it allows one to pursue riskier research agendas (see Alchian 1977). Moreover, we have ignored moral hazard, hierarchies, and other interactions across workers. It may also be said that frequently when one is denied tenure, he moves to a position that is not qualitatively different from the previous position. This is not the case in our model. Yet we believe that our job-matching framework answers a basic question about tenure, namely, why do people with relatively poor records leave instead of simply having their salary reduced? The answer must surely be that the poor record is evidence that the worker is more productive elsewhere.

[3] When workers are risk neutral and their subjective discount rate equals the market interest rate, they are indifferent not only among distributions of consumption across states but also across periods. In this case, access to the capital market is irrelevant. When workers are risk averse, i.e., their period utility function is concave, access to the capital market matters.

decision to the firm (subject to the constraint that the decision maximize the ex ante expected utility of the worker). For certain realizations of the output process, termination of employment will not be voluntary on the part of the worker. That is, while it is optimal to agree ex ante to leave in certain situations in order to receive a higher current wage, leaving may not be ex post optimal for the worker once the situation is realized. It is in this sense that the model with risk-averse workers exhibits involuntary separations.

A worker in this model obtains tenure if for any future realization of the output sequence the optimal strategy is to retain him. We again show that a tenure standard r^* exists; that is, a worker who achieves r^* is assured of being retained thereafter. An up-or-out age T also exists in the sense that employment is never terminated later than that age. Contrary to the case of risk neutrality, however, some workers may achieve tenure without attaining r^* by age T. Finally, we consider some efficiency implications of the model with risk-averse workers. We show that workers may stay in their current job even when it would be socially optimal to switch. The intuition for this has to do with risk aversion and is spelled out in Section III.

II. A Model with Risk-Neutral Workers

Model Description

The model is composed of many employers and many workers. There are two mutually exclusive jobs. These may be thought of as being in the same or different firms.

A worker in the primary job is assumed to produce either zero or one unit of output at each age, which the firm can then market at a fixed price (which we normalize to one). All workers are assumed to have a finite working life of length A. Workers are homogeneous *except* that some have a high expected output per unit time q_H at any given age while others have a low expected output per unit time q_L at any given age. Naturally, we assume $0 \leq q_L < q_H$. The output in the primary job of a worker of productivity $q \in \{q_L, q_H\}$ is assumed to follow a Poisson process with parameter q.

In a Poisson process, time (really time in process or "age" in this model) is viewed as continuous. At any age $a \in [0, A]$, the probability that a worker of type q will have produced a cumulative output of $R_a = r$ units is $\Pr(R_a = r|q) = e^{-qa}(qa)^r/r!$ for $r = 0, 1, \ldots$. As mentioned above, q may be interpreted as the expected output per unit of time.

No one, including the worker himself, knows the productivity of any given worker. A priori, each worker is believed by everyone to have low productivity with probability π independently of any other worker ($0 < \pi < 1$). Beliefs about a given worker of age $a > 0$ are assumed to be formed according to Bayes's rule, given the worker's age, a, and his pattern of outputs at ages prior to a. It is easy to show that cumulative output r (for "record") is a sufficient statistic for the complete pattern of previous outputs (see, e.g., DeGroot 1970, exercise 1, pp. 183–84). Thus the posterior probability that any given worker of characteristics (a, r) (i.e., age a, record r) has low productivity, q_L, is

$$\pi(a, r) = \Pr(q = q_L | R_a = r) = \frac{\pi e^{-q_L a} q_L^r}{\pi e^{-q_L a} q_L^r + (1 - \pi) e^{-q_H a} q_H^r} \quad (1)$$

for any $(a, r) \in C = [0, A] \times Z$ (Z is the set of natural numbers, 0, 1, 2, . . .). We will call C the set of feasible characteristics of workers. Thus the expected productivity of a worker $(a, r) \in C$ is given by

$$q(a, r) = q_L \pi(a, r) + q_H[1 - \pi(a, r)]. \quad (2)$$

From (1), it follows that π is strictly increasing in a for given r and strictly decreasing in r for given a. Thus from (2), q is strictly decreasing in a for given r and strictly increasing in r for given a.[1]

Firms are assumed to be risk neutral and the labor market is assumed to be competitive. This implies that, in equilibrium, firms will make zero expected profits. Since in this model workers are also risk neutral, we can assume, without loss of generality, that wages in the primary job are equal to perceived expected productivities
Productivity in the secondary job is assumed to
tainty to be W per unit of time for every work
and any record he may have in the primary
job can be assumed, without loss of generality
salary of W per unit of time until retirement
worker must decide only whether to select the
job at that age. Switching from the primary job
any age $a > 0$ is assumed to involve a once-and-for-all
A worker of age $a = 0$ may choose either job at no cost. Switching from the secondary job to the primary job is assumed to be free, however, it is easy to show that no worker will take advantage of this option. First suppose that there are no moving costs. In deciding

---

[1] These statements are correct only if $q_L > 0$. If $q_L = 0$, then $\pi(a, 0)$ and $q(a, 0)$ have the properties mentioned, but $\pi(a, r) = 0$ and $q(a, r) = q_H$ for any a and $r \geq 1$.

[2] The term "moving" cost should not be taken too literally. The cost K is meant to reflect all sorts of unreimbursed costs, monetary and psychological, occasioned by the switch in career. To the extent that direct moving expenses are reimbursed by the new employer in lieu of higher wages, these are included in K. In a more general model, one might assume that such costs increase with age or tenure in the firm. Letting K increase with a in the model would not affect the qualitative results but would tend to reduce the equilibrium up-or-out age.

whether to choose the primary or secondary job at any age, a worker must trade off the benefits to gaining additional information about his productivity in the primary job by pursuing it over the next small time interval against the additional current wage he may get in the secondary job (of course, if the wage in the secondary job is smaller than the current wage in the primary job, he will not choose the former at that age). Suppose a worker chooses the secondary job at age a when his record in the primary job is r. Consider his situation an instant later, assuming he did not switch back in the interim. His perceived productivity in the primary job is the same as when he chose the secondary job at age a since nothing is revealed while in the secondary job. The secondary wage is also the same. Moreover, the additional information about the primary job that he can get now is less valuable because there is less time left in which to capitalize on that information. Therefore, the primary job looks worse relative to the secondary job than it did when he chose the secondary job. Thus, even in the absence of costs of moving from the primary job to the secondary job, he will not now choose the former.[6] Moving costs simply reinforce this argument; that is, the secondary job looks even more attractive now since he can choose it without paying the moving cost, K, which he had to pay when he chose it before. Consequently, the worker will not switch back to the primary job once he has chosen the secondary job.

Workers take the salaries offered in the primary job, $q(a, r)$, and in the secondary job, W, as given and make an expected wealth-maximizing career choice at each age. All workers are assumed to face the same, constant market real interest rate ρ per period. Let $V(a, r)$ be the expected wealth (present value of all future income) of a worker (a, r) assuming that he makes an optimal career choice at each future age. Consider a very short time interval h and suppose that the worker (a, r) decides to stay in the primary job during the period a to $a + h$. Then, to a close approximation, he will produce zero units of output in that interval with probability $1 - q(a, r)h$ and one unit with probability $q(a, r)h$ and more than one unit with probability zero. This statement becomes exact in the limit as h approaches zero.

It follows, therefore, that, to a close approximation, the expected wealth from staying in the primary job at (a, r) from a to $a + h$ is (for small h)

$$\frac{q(a, r)h + q(a, r)hV(a + h, r + 1) + [1 - q(a, r)h]V(a + h, r)}{1 + \rho h}. \tag{3}$$

---

[6] This argument is adapted from Rothschild (1974), who discussed the discrete-time, infinite-horizon case. A formal proof for the continuous-time, finite-horizon case is available from the authors on request.

In this equation, $q(a, r)h$ is the salary over the period from a to $a + h$. It is assumed that the salary is paid at $a + h$, although as h gets small this is irrelevant. The expected (and actual) wealth from switching to the secondary job at (a, r) is

$$v(a) = W\xi(a) - K \tag{4}$$

for $a > 0$ and $v(0) = W\xi(0)$, where $\xi(a)$ is the present value of a continuous flow of \$1.00 per period over the interval $[a, A]$ using interest rate ρ, that is,

$$\xi(a) = \left(\frac{1}{\rho}\right)[1 - e^{-\rho(A-a)}] \tag{5}$$

Since the worker can choose to stay with the firm for an interval $(a, a + h)$ or to quit, and since he will never switch from the secondary to the primary job, the expected wealth of worker (a, r) is either $v(a)$ as defined in (4) or the expression in (3), whichever is larger, that is,

$$V(a, r) \cong \max \left\{ v(a), \frac{q(a, r)h + q(a, r)hV(a + h, r + 1) + [1 - q(a, r)h]V(a + h, r)}{1 + \rho h} \right\} \tag{6}$$

If $v(a)$ is the larger of the two expressions for small h, then $V(a, r) = v(a)$. If, however, $v(a)$ is the smaller of the two, then multiplying by $1 + \rho h$ and subtracting $V(a, r)$ from both sides gives

$$\begin{aligned}\rho h V(a, r) \simeq\ & q(a, r)h + q(a, r)hV(a + h, r + 1) \\ & + [1 - q(a, r)h][V(a + h, r) - V(\end{aligned}$$

Now dividing by h and taking limits as h appr

$$V(a, r) = \left(\frac{1}{\rho}\right)\{q(a, r) + q(a, r)[V(a, r + 1)$$

where dot over a function denotes its (partial) derivative [illegible] to age.

Therefore, we have

$$V(a, r) = \max \left\{ v(a), \left(\frac{1}{\rho}\right)\{q(a, r) + q(a, r)[V(a, r + 1) - V(a, r)] + \dot{V}(a, r)\} \right\}. \tag{FE}$$

The above equation (FE) is the functional equation for the dynamic programming problem faced by any worker. In order to define the problem fully, however, we must specify a boundary condition. The obvious boundary condition for this problem is that the present value of future income starting at retirement age A is zero for any record, that is,

$$V(A, r) = 0 \quad \text{for all } r. \tag{B}$$

The two equations (FE) and (B) define the value function V, if a solution of (FE) that satisfies (B) exists (a solution is constructed below).

Any worker (a, r) will choose to exit from the primary job if and only if

$$\rho v(a) > q(a, r) + q(a, r)[V(a, r + 1) - V(a, r)] + \dot{V}(a, r). \tag{7}$$

We define the exit set, E, as the set of characteristics that will result in exiting the primary job, that is, $E = \{(a, r) \in C \mid (7) \text{ holds}\}$. Note that E is fully determined by ρ, W, and K.

The analysis of the model consists of solving the optimal stopping problem defined by (FE) and (B) in order to characterize the exit set E. The solution will involve an increasing sequence of stopping times, $t(0) < t(1) < \ldots$ such that (a, r) is in E if and only if $a > t(r)$. This analysis is carried out in the Appendix; here we present the main conclusion informally.

Informal Description of the Main Results

The nature of the solution to the problem stated above can be described informally with the aid of figure 1. All the turnover characteristics of the model derive from the shape of the boundary of the exit set. This boundary is the set of points $[t(r), r]$ shown in figure 1 as heavy dots.

The boundary of the exit set must be a nondecreasing function of age. The reason is that the expected output per unit of time of a worker of a given age is an increasing function of record, that is, $q(a, r) < q(a, r')$ for $r' > r$. A worker who has attained a given record r' by age a is more likely to be of high quality than one who has attained only r. Thus if a worker with record r chooses to leave at age $t(r)$, a worker with the better record r' will not leave before $t(r)$; that is, $t(r)$ is increasing in r (see lemmas 2 and 7 in the App. for a formal treatment).

The boundary of the exit set cannot touch the vertical lines through $a = 0$ or $a = A$; that is, $t(0)$ cannot be at $a = 0$ nor can $T = t(r^* - 1)$ be at $a = A$. If one assumes that at time zero the typical worker (with record $r = 0$) weakly prefers the primary job, then, due to positive mobility costs, he will stay for a while even if no output is forthcoming. Similarly, due to finite working life and positive costs of mobility, if the worker leaves, he will always do so strictly before retirement

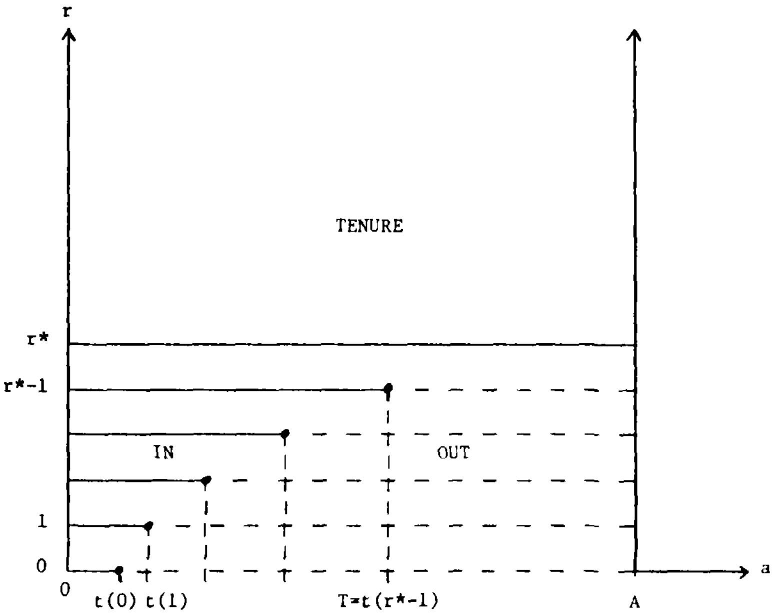


FIG. 1 — Properties of the solution

age. Indeed, at ages close to A, moving costs exceed the present value of the salary in the secondary job.

Finally, there is a record r^* sufficiently high [illegible]
who attains it will never leave. The boundary [illegible]
cross the horizontal line through r^*. This, again [illegible]
horizon and the existence of moving costs. [illegible]
estimates of productivity, equations (1) and (2) [illegible]
imply that the downward revision in the work[illegible]
tivity is bounded; that is, the worst possible futu[illegible]
mate for a worker (a, r) is $q(A, r)$. If this lower bou[illegible]
wage in the secondary job, W, minus the annual equivalent of moving costs, $K/\xi(a)$, the worker will clearly stay in the primary job until retirement (see theorems 1 and 2).

The considerations above imply an exit set of the form drawn in figure 1. Workers with characteristics in the region labeled "OUT" in figure 1 will not be observed in the primary job because, when the boundary is crossed, the worker switches to the secondary job. Workers with records better than r^* will never be observed to leave the primary job; that is, their probability of future exit is zero. These workers have "tenure." Finally, workers with characteristics in the region labeled "IN" may be viewed as experimenting on the job.

Their probability of future exit is strictly between zero and unity. An interesting aspect of the gradual sorting process is that after some critical age (labeled T in fig. 1), experimentation, and thus mobility, will cease. All workers with a record equal to or better than r^* will choose to stay in the primary job until retirement; everyone else will leave. This follows from the finite horizon and the fact that realizations of the Poisson process are nondecreasing step functions in age-record space. The employment status of the worker can change only in the range in which the boundary of the exit set is strictly increasing.

The turnover pattern just described corresponds quite well with observed mobility patterns. In academia, for instance, where output (publications) is publicly observable, workers are required to attain a minimal standard (i.e., number of quality-adjusted publications) to obtain tenure. Those who have not met the tenure standard by a certain professional age seek employment at another institution, usually of lower quality or reputation or of a nonacademic nature. Some workers (those with poor records) move before the up-or-out age. Similar patterns are observed in other occupations.

Computation and Comparative Dynamics

In this subsection we discuss computation of the tenure standard, r^*, and up-or-out age, T (a technique for calculating the rest of the solution is given in the App.). We also present some comparative statics results for these quantities and a numerical example. The example shows that, in fact, this model is capable of producing values of r^* and T that are similar to observed values using values of exogenous parameters that are also fairly realistic.

To calculate r^*, note that if a worker plans to remain in the primary job until retirement, his expected wealth is simply the present value of a constant stream equal to his *current* productivity at each age, namely, $q(a, r)\xi(a)$. This follows from the fact that the expected value of one's future perceived productivity equals his current perceived productivity; that is, perceived productivity q is a Martingale. Thus, for all $r \geqq r^*$, $V(a, r) = q(a, r)\xi(a) \geqq v(a)$, and r^* is the minimum such value of r; that is, r^* is the smallest integer r such that[7]

$$q(a, r) \geqq W' - \frac{K}{\xi(a)} \quad \text{for all } a. \tag{8}$$

Any change that increases the right-hand side of (8) will tend to increase r^* (although very small changes will have no effect on r^* because of the integer restriction). Since $\xi(a)$ increases, for any a, with an increase in A or a decrease in ρ, the tenure standard, r^*, will tend to increase with increases in the secondary wage, W, increases in retirement age, A, decreases in moving cost, K, and decreases in the real interest rate, ρ.

---

[7] This is proved formally as lemma 8 in the App

Intuitively, any change that either makes the secondary job more attractive relative to the primary job or reduces the annualized equivalent of moving costs, $K/\xi(a)$, will induce people with a record equal to the initial tenure standard to leave. Thus a better record is now required for it to be optimal to stay in the primary job. This same reasoning indicates that any change that makes the primary job more attractive, that is, increases $q(a, r)$ for any (a, r), will have the opposite effect on r^*. Since q increases with increases in q_H, q_L, and $1 - \pi$ for any (a, r), the tenure standard, r^*, will tend to fall with increases in the productivity q_H or q_L, of either type of worker, or with increases in the fraction of workers who are highly productive, $1 - \pi$.

Comparative dynamics results on the up-or-out age, T, are more difficult to obtain. First recall that the up-or-out age $T = t(r^* - 1)$. Using lemma 7 (see the App.) and the fact that, for $r \geqq r^*$, $V(a, r) = q(a, r)\xi(a)$, T is the smallest age a at which

$$\begin{aligned} 0 = {} & W[1 + q(a, r^* - 1)\xi(a)] - K[\rho + q(a, r^* - 1)] \\ & - q(a, r^* - 1)[1 + q(a, r^*)\xi(a)]. \end{aligned} \tag{9}$$

Since it can be shown that the right-hand side of (9) is increasing in a at $a = T$, we see that T varies inversely with W and di[illegible] provided that the changes in W or K are small en[illegible] Although it does not seem possible to say, in ge[illegible] T when the changes in W or K also change r^*, s[illegible] ples suggest that increases in W will increase [illegible] increases in K have the opposite effect.

Finally, we would like to point out that it is [illegible] examples with realistic values of T and r^* using [illegible] the exogenous parameters. For instance, if the prior [illegible] productivity is that a worker has a 50-50 chance of having productivity $q_L = 1$ or $q_H = 2.8$ units per period, the market interest rate is $\rho = 10$ percent per period, the retirement age is $A = 40$ periods after entry, the secondary wage is $W = 1.1$, and moving costs are $K = 0.94$ (i.e., 85 percent of W), then the tenure standard $r^* = 8$ units and the up-or-out age T is between 8 and 9 periods.

III. Risk Aversion and Imperfect Capital Markets

The analysis of Section II considered risk-neutral workers and firms and concluded that competition will force an equality between wages and the perceived productivity of the workers at any instant. There is thus no insurance in that model. Yet it would appear that the insurance features of observed contracts are important (especially where tenure is involved). Therefore, in this section, we discuss the extent to which worker risk aversion and lack of access to capital markets coupled with the model of information evolution and job matching of Section II can account for both the turnover and insurance aspects of observed contracts.[8]

The introduction of worker risk aversion has a major effect on the form of an optimal wage contract. While under risk neutrality wages fluctuate with productivity, wage insurance is provided by the firm when the worker is risk averse. This takes the form of a downward rigid wage profile; that is, a worker's wage will never decline while he is employed in the primary job and will increase only when necessary to prevent the worker from moving to a primary job in another firm.[9]

Despite this difference in the wage contract, there are some similarities in the turnover pattern. Let us define a tenure standard r^* to be a critical value of r such that for all $r \geqslant r^*$, a worker who attains r at some age a is guaranteed to be employed by the firm throughout the remainder of his working life. This standard is defined so as to apply for any history. It can be shown that there exists a positive probability of reaching r^* following the optimal contract.[10] This follows from the observation that a worker whose perceived productivity at any given age exceeds his opportunity cost (i.e., his productivity in the secondary job) will not be rejected at that age. Similarly we may define the up-or-out age as the latest age, T, such that for all ages $a \geqslant T$, employment in the primary job is guaranteed irrespective of output. For $a < T$, there is still some experimentation, and the continuation of employment in the primary job depends on performance. It can be shown that such a critical age exists.[11] This follows from the fact that it is relatively expensive for a worker to move to the secondary job late in his career, and the gain from acquiring information by continuing the experimentation period is declining. The considerations above are valid whether workers are risk averse or risk neutral, and therefore a tenure standard and up-or-out age exist in both cases. Contrary to the case of risk neutrality, however, the tenure standard and up-or-out age are not tied together in the sense that, under risk aversion, some workers with records less than r^* will still be in the primary job after age T. (This is similar to the results one gets from Jovanovic's [1979] model.) The reason for this is that with risk aversion, the tenure decision depends also on the wage commitments that the worker has accumulated from the past as part of the insurance contract. Those commitments may be different for workers of the same age and record because of differences in history.

---

[8] As mentioned in the Introduction, our model in this section is an extension of those of Freeman (1977), Weiss (1981), and Harris and Holmstrom (1982). The essential differences were outlined in the Introduction

[9] As in Harris and Holmstrom (1982), if workers could commit themselves to remaining with a particular firm, wages would be completely rigid.

[10] For proof of this statement, the reader is referred to our working paper, Harris and Weiss (1982).

[11] For proof of this statement, the reader is referred to Harris and Weiss (1982).

To illustrate this last point and gain some additional insight into the wage process under risk aversion, consider two workers, Fast and Slow. Although they are the same age a and have the same cumulative output r, and therefore the same perceived productivity q, Fast produced his r units in his first r periods while Slow produced his in his last r periods. Since Fast had r successes in his first r periods, his wage increased rapidly to a level higher than Slow's wage. Slow's wage, in fact, would still be well below this level even after achieving his r successes because these occurred over $a > r$ periods. Thus, because of the downward rigidity of Fast's wage, even though by age a both have the same record and perceived productivity, Fast has a higher current wage and promised future wages than does Slow. This can result in it being optimal for Fast to remain in the primary job until retirement even if he never produces another unit of output and for Slow to leave before retirement. In this case, if r^* is defined as in the previous paragraph, we see that there is no $T < A$ such that all workers with records worse than r^* at age T leave ($r < r^*$ since Slow does not achieve tenure, but Fast always has record r and still achieves tenure. The point here is that a worker's current situat
terized simply by his current age and cumulat
case.

An important difference between the risk-
cases is that in the latter wage rigidity is a sou
inefficiency. To see this, we introduce the fo
sider a worker at age a and let $X(a)$ be the
expected net costs (i.e., wages minus output) to the
remainder of the worker's life (i.e., from a to A). One may also think of $X(a)$ as the net surplus (above-average product) promised by the firm to the worker. Clearly the firm is indifferent among all rearrangements of wages and employment that keep $X(a)$ fixed since the firm is assumed to be risk neutral and to have access to perfect capital markets. Define $V(x, a, r)$ as the maximal expected utility that a worker of age a can obtain given that $X(a) = x$ and given that his success record is r, where the maximization is over all feasible future wage and employment paths. Also let $v(a)$ be the present value of the utility of the constant wage stream in the secondary job from age a to age A minus the moving cost (as in Sec. II). A necessary condition for

ex post social optimality is that a worker of age a and record r leave the primary job whenever the value of staying with zero expected cost to the firm ($x = 0$) is less than the value of moving. This can be stated mathematically for the two cases corresponding to success or failure (to produce) at age a as

$$V(0, a, r + i) < v(a), \quad \text{for } i = 0, 1. \tag{10}$$

Under an optimal wage and employment contract, the decision whether to retain the worker depends on the comparison of the *joint* benefit and cost to the worker and the firm of retaining the worker. The joint benefit can be expressed mathematically for the two cases of success and failure at age a as

$$V(x + \Delta_i, a, r + i) - (x + \Delta_i)V_x(x, a, r), \quad i = 0, 1,$$

where Δ_1 is the amount by which the firm's optimal commitment changes if a jump in output occurs at age a, $\Delta_0 = 0$, and V_x is the marginal value to the worker of an increase in the level of commitment. In this expression, the benefit to the firm of retention, $-(x + \Delta_i)$, is multiplied by the marginal value of this commitment to the worker because V_x is the marginal rate of transformation between the utilities of the firm and worker (since the firm is risk neutral). The joint cost of retention is simply the opportunity cost of not moving to the secondary job, namely $v(a)$. Thus the conditions for *not* retaining the worker are

$$V(x + \Delta_i, a, r + i) - (x + \Delta_i)V_x(x, a, r) < v(a), \quad i = 0, 1. \tag{11}$$

Condition (11) cannot hold unless $x + \Delta_i > 0$, that is, unless the worker imposes a net cost on the firm. In this case, there is no need to raise the worker's wage to prevent him from moving to another firm in the primary sector, and it follows (see the App.) that

$$V_x(x, a, r) = V_x(x + \Delta_i, a, r + i), \quad i = 0, 1.$$

Thus, if V is sufficiently concave in x locally, it is possible to have

$$V(0, a, r + i) < v(a) \leq V(x + \Delta_i, a, r + i) - (x + \Delta_i)V_x(x, a, r),$$

in which case the worker will stay even when it is ex post socially inefficient to do so. The local concavity of V in x will be inherited from the concavity of u, which in turn implies both risk aversion and a desire for a smooth wage profile over time (if the worker has no access to the capital market). The inefficiency arises because the provision of wage insurance and banking services is tied to the worker's employment. If these aspects can be separated, for instance, by severance payments and/or provision of an annuity to a worker who agrees to quit, then a privately optimal employment strategy will also be ex post efficient.

An additional important difference from the case of risk neutrality is that, under risk aversion, separations are not generally voluntary. As can be seen from condition (11), it may be optimal to expel the worker even though the value to the worker of remaining in the primary job exceeds the value of the secondary job. This will happen if the cost of the firm's insurance commitment to the worker outweighs the net benefits to the worker of remaining in the primary job. Under the optimal contract, it may turn out that, after discovering that his situation calls for leaving the primary job, the worker would prefer to continue with a contract that is optimal thereafter but stipulates retention currently. To avoid such inconsistencies, it is optimal ex ante to delegate to the firm the power to fire the worker when the ex ante optimal contract specifies moving to the secondary job. It is in this sense that, under risk aversion, not all separations are voluntary.

IV. Conclusions

We have explored the consequences of introducing workers with finite horizon and risk aversion into a simple job-matching framework. We show that the finite-horizon assumption alone can account for several aspects of observed turnover patterns that are not consistent with similar infinite-horizon models. In particular, the model generates an up-or-out age, T, and a tenure standard, [illegible] such that all workers will stop experimenting with
productivity is unknown in favor of one in wh
known at or before age T unless their cumu
ceeds the tenure standard by that age. In the
will remain in the uncertain job until retireme

When risk aversion on the part of the worke
rium involves contracts in which wage insuran
is provided. This results in the possibility of involu
and ex post inefficiency of the employment pattern. Although there will also be a tenure standard and up-or-out age in this model, generally these will not have the property that all workers who have not achieved tenure by the up-or-out age move to the certain job.

Fruitful extensions of the models would be to introduce various types of worker activities within the uncertain job. For example, introducing unobservable effort would allow one to analyze the extent to which insurance aspects of optimal contracts lead to excessive shirking in this dynamic, job-matching framework. Introducing an investment decision would allow consideration of the effect of contracts on the incentives to undertake riskier projects.

Appendix

The Risk-Neutral Model

DEFINITION: A path to a characteristic (a, r) is any feasible sample path of a Poisson process over the interval $[0, a]$.

Let $P(a, r)$ be the set of paths to (a, r) and $E'(a, r)$ be the set of paths to (a, r) that do *not* pass through E.

LEMMA 1: $V(a, r)$ in (7) can be replaced by $v(a)$; that is, for any (a, r), (7) holds at (a, r) if and only if

$$\rho v(a) > q(a, r) + q(a, r)[V(a, r + 1) - v(a)] + \dot{v}(a). \qquad \text{(A1)}$$

PROOF: First suppose (7) holds at (a, r). Then, from (FE), $V(\cdot, r) \equiv v(\cdot)$ in a neighborhood of a. Therefore $\dot{V}(a, r) = \dot{v}(a)$ and (A1) follows by substitution.

Now suppose (A1) holds. Then (A1) holds on a neighborhood of a, so v satisfies (FE) on a neighborhood of a. This implies that $V(\cdot, r) \equiv v(\cdot)$ on this neighborhood, for suppose not. Then

$$\begin{aligned}\rho V(a, r) &= q(a, r) + q(a, r)[V(a, r + 1) - V(a, r)] + \dot{V}(a, r)\\ &\geq \rho v(a) \quad \text{from (FE)}\\ &> q(a, r) + q(a, r)[V(a, r + 1) - v(a)] + \dot{v}(a) \quad \text{from (A1)},\end{aligned}$$

so

$$0 \geq q(a, r)[v(a) - V(a, r)] > \dot{v}(a) - \dot{V}(a, r)$$

or

$$\dot{v}(a) > \dot{V}(a, r).$$

But from (A1),

$$\dot{v}(a) < [\rho + q(a, r)]v(a) - q(a, r)[1 + V(a, r + 1)],$$

and from (FE),

$$\dot{V}(a, r) = [\rho + q(a, r)]V(a, r) - q(a, r)[1 + V(a, r + 1)].$$

Therefore $v(a) > V(a, r)$, which contradicts (FE). We conclude that $V(\cdot, r) \equiv v(\cdot)$ on this neighborhood of a, and the result follows from substitution. Q.E.D.

LEMMA 2: If a worker leaves the primary job, then any worker of the same age but inferior record will also leave; that is, if $(a, r) \in E$ and $r' \leq r$, then $(a, r') \in E$.

PROOF: By backward induction on r. Clearly the lemma holds for $r' = r$. Suppose it holds for all $r \geq r' \geq s$ for some $s < r$. We will show it holds for $r' = s - 1$. By lemma 1, it suffices to verify that (A1) holds at $(a, s - 1)$. Now by the induction hypothesis and lemma 1, (A1) holds for all $r \geq r' \geq s$. Therefore $V(a, r) = v(a)$ for all $r \geq r' \geq s$ and

$$\rho v(a) > q(a, s) + q(a, s)[V(a, s + 1) - v(a)] + \dot{v}(a)$$

or

$$\rho v(a) > q(a, s - 1) + q(a, s - 1)[V(a, s) - v(a)] + \dot{v}(a)$$

since $q(a, s - 1) < q(a, s)$ and $V(a, s) = V(a, s + 1) = v(a)$. This last inequality is just (A1) at $(a, s - 1)$. Q.E.D.

Lemma 3: If there is some way of reaching (A, r) without exiting, then no one with a record of r or better will ever exit; that is, if r is such that $E'(A, r)$ is nonempty, then for any $s \geq r$ and any a, (a, s) is not in E.

Proof: Suppose (a, s) is such that $s \geq r$ and $(a, s) \in E$. Now, for any $p \in P(A, r)$, $p(a) \leq r \leq s$ since p is nondecreasing. Therefore, by lemma 2, $[a, p(a)] \in E$. Since this is true for any $p \in P(A, r)$, $E'(A, r)$ is empty contrary to the assumption. Q.E.D.

Lemma 4: If $E'(A, r)$ is nonempty, then for any a and $s \geq r$, $V(a, s) = q(a, s)\xi(a)$.

Proof: First we must establish a certain property of q. Using equations (1) and (2), it is easy to check that, for any (a, r),

$$\dot{q}(a, r) = -q(a, r)[q(a, r + 1) - q(a, r)]. \tag{A2}$$

Since $\xi(A) = 0$, $q(a, s)\xi(a)$ satisfies (B). We need only show that it satisfies (FE), that is, for $s \geq r$,

$$\rho V(a, s) = q(a, s) + q(a, s)[V(a, s + 1) - V(a, s)] + \dot{V}(a, s).$$

But substituting $q(a, s)\xi(a)$ for $V(a, s)$ and $q(a, s + 1)\xi(a)$ for $V(a, s + 1)$ gives

$$\begin{aligned}\rho q(a, s)\xi(a) &= q(a, s) + q(a, s)[q(a, s + 1) - q(a, s)]\xi(a)\\ &\quad + [\dot{q}(a, s)\xi(a) + q(a, s)\dot{\xi}(a)]\\ &= q(a, s)[1 + \dot{\xi}(a)], \quad \text{using (A2)},\\ &= q(a, s)\rho\xi(a), \quad \text{since } \dot{\xi}(a) = \rho\xi(a) - 1\end{aligned}$$

Thus $q(a, s)\xi(a)$ does satisfy (FE) Q.E.D.

Definition: An age $T > 0$ and record r^* is called a tenure characteristic if

(i) No worker ever exits the primary job after achieving a record of r^* or better.

(ii) If $r^* > 0$, $T < A$ and no worker ever remains in the [illegible] reaching T with record $r^* - 1$ or worse. If $r^* = 0$, T [illegible]

(iii) If $r^* > 0$, there exists a path p to $(T, r^*$ [illegible] through the exit set and such that $p(a) = r^* - 1$ [illegible]

The age T is called the up-or-out age and the rec[illegible] standard. The third condition guarantees that no[illegible] worse than r^* exit before the up-or-out age

Theorem 1. Suppose for some r, $E'(A, r)$ is no[illegible]

$$\begin{aligned} r^* &= \min\{r \in Z \mid E'(A, r) \text{ is nonempty}\},\\ X &= \{a \in [0, A] \mid (a, r^* - 1) \text{ is not in } E\} \quad \text{if } r^* > 0,\\ &= [0, A] \quad \text{if } r^* = 0,\\ T &= \sup X.\end{aligned}$$

Then (T, r^*) is the unique tenure characteristic and $X = [0, T)$ if $T < A$, $X = [0, A]$ if $T = A$.

Proof: We must first show that X is nonempty for $r^* > 0$. Let $p \in E'(A, r^*)$. Then for some age b, $p(b) = r^* - 1$. Therefore, $(b, r^* - 1)$ is not in E and $b \in X$.

That (T, r^*) satisfies part i of the definition of a tenure characteristic is obvious from lemma 3 and the definition of r^*.

To prove part iii, suppose $r^* > 0$ and let $p \in E'(A, r^*)$. Also let b be such that $p(b) = r^* - 1$. Then $b \in X$ so $b \leq T$. If $b < T$, define p' by $p' = p$ on $[0, b]$

and $p' \equiv r^* - 1$ on $[b, T]$. If $b = T$, then there is an age $c < b$ such that $p(a) = r^* - 2$ on $[c, b)$. Choose $c < d < b$ and define $p' \equiv p$ on $[0, d)$, $p' \equiv r^* - 1$ on $[d, T]$. Then p' is the required path.

Now suppose $r^* > 0$ and $T = A$ Then by the construction above $E'(A, r^* - 1)$ is none[illegible] the definition of r^*. Therefore $T < A$. Part ii now follows [illegible] definition of T and lemma 2.

Next we show that $X = [0, T)$ if $T < A$. Clearly, by definition of T, X is contained in $[0, T)$. To show the reverse inclusion, let a be any element of $[0, T)$, and let $p \in E'(T, r^* - 1)$ as constructed above. Then $[a, p(a)]$ is not in E and $p(a) < r^*$. Therefore by lemma 2, $(a, r^* - 1)$ is not in E. Hence $[0, T)$ is contained in X If $T = A$, it is obvious that $X = [0, A]$.

Finally suppose (T', r') is another tenure characteristic. Then $r' \geq r^*$. Suppose $T' < T$ Let $p \in E'(T, r^* - 1)$. Then for $a \in (T', T]$, $[a, p(a)]$ is not in E, which contradicts part ii of the definition since $r^* \leq r'$ Therefore $T' \geq T$. Now suppose $r' > r^*$. Then by lemma 3, $T' = A$, which violates part ii of the definition. Therefore $r' = r^*$. Finally, since $r' = r^*$, the argument that $T' \geq T$ can be reversed to yield $T \geq T'$. Thus $T = T'$. Q.E.D.

Assumption: Workers (weakly) prefer to start out in the primary job; that is, (A1) does not hold at $a = r = 0$.

Lemma 5. (a) $q_H > W$ and (b) $\lim_{h \to 0} V(h, 0) = V(0, 0)$; that is, $V(\cdot, 0)$ is continuous at 0.

Proof: Part a is obvious since otherwise the wage in the secondary job would exceed the maximum possible wage in the primary job, and no one would weakly prefer the latter at (0, 0)

Now for $h > 0$, but small, the value of choosing the primary job at (0, 0) is approximately

$$q(0, 0)h + \frac{q(0, 0)hV(h, 1) + [1 - q(0, 0)h]V(h, 0)}{1 + \rho h}.$$

The relation is exact in the limit as h approaches zero. Therefore, since the primary job is weakly preferred at (0, 0), the limit of the expression above equals $V(0, 0)$ But that limit is $\lim_{h \to 0} V(h, 0)$. Q.E.D.

Lemma 6: For some r, $E'(A, r)$ is nonempty; that is, there is a path to (A, r) that does not pass through E

Proof: Since $q_H > W$ and $\lim_{r \to \infty} q(A, r) = q_H$, for r sufficiently large, $q(A, r) > W$. Let $\bar{r}$ be any such r Then $q(a, \bar{r}) > W$ for any $a \geq 0$. It follows that $(a, \bar{r}) \notin E$ for every $a > 0$. Now suppose $p \in E'(a, \bar{r})$ for some $a > 0$. Then define p' by $p' \equiv p$ on $[0, a]$, $p' \equiv \bar{r}$ on $[a, A]$. In this case $p' \in E'(A, \bar{r})$, and we are done. Thus it suffices to show that for some $a > 0$, $E'(a, \bar{r}) \neq \phi$. Suppose that for every $a > 0$, $E'(a, \bar{r}) = \phi$. Then for any $a > 0$ there is a $0 < t < a$ such that $(t, 0) \in E$, that is, $v(t) = V(t, 0)$ This implies (using lemma 5) that $v(0) - K = \lim_{a \to 0} v(a) = \lim_{a \to 0} V(a, 0) = V(0, 0)$ or, since $K > 0$, $v(0) > v(0) - K = V(0, 0)$, which is a contradiction. Q.E.D.

Theorem 2: (T, r^*) defined in theorem 1 is the unique equilibrium tenure characteristic.

Proof: By lemma 6, there is an r such that $E'(A, r)$ is nonempty. Therefore r^* and T are well defined. The theorem now follows from theorem 1. Q.E.D.

Lemma 7: Suppose that $r^* > 0$ is an equilibrium tenure standard. Define $X(r) = \{a | (a, r) \text{ is not in } E\}$, and $t(r) = \sup X(r)$ for $r = 0, \ldots, r^* - 1$. Then $t(r)$ is nondecreasing in r and $X(r) = [0, t(r))$.

Proof: Suppose $r' > r$ and $t(r') < t(r)$. By definition of t, there is $t(r') < a < t(r)$ such that (a, r) is not in E but $(a, r') \in E$. Since $r' > r$, this contradicts lemma 2. The remainder of the lemma is proved exactly as in the proof of theorem 1. Q.E.D.

Computation of the Solution

Lemma 8: A worker remains in the primary job until retirement if and only if the present value of a constant income equal to his current wage until retirement exceeds the value of switching; that is, $E'(A, r)$ is nonempty if and only if $q(a, r)\xi(a) \geq v(a)$ for all a.

Proof: If $E'(A, r) \neq \emptyset$, then by lemma 4, $V(a, r) = q(a, r)\xi(a)$. From (FE), $V(a, r) \geq v(a)$ for all a, so $q(a, r)\xi(a) \geq v(a)$ for all a.

Now suppose $q(a, r)\xi(a) \geq v(a)$ for all a. Then this inequality holds for all $r' \geq r$. It is easy to check that $V(a, r') = q(a, r')\xi(a)$ for all $r' \geq r$ satisfies (FE) and (B) for all $r' \geq r$. Therefore, $(a, r) \notin E$ for all a. It can now be shown, as in the proof of lemma 6, that $E'(A, r) \neq \emptyset$. Q.E.D.

From lemma 8 we can calculate r^* as the smallest $r \in Z$ such that $q(a, r)\xi(a) \geq v(a)$ for every a. We can now construct V by backward recursion on r starting with r^*. If $r^* = 0$, let $V(a, r) = q(a, r)\xi(a)$ for all (a, r). Otherwise let

$$H(a, r) = \rho v(a) - q(a, r) - q(a, r)[V(a, r + 1) - v(a)] - \dot{v}(a)$$

Now for $r \geq r^*$, let $V(a, r) = q(a, r)\xi(a)$.

Define $X(r)$ and $t(r)$ as in lemma 7. Then from lemmas 1 and 7, $t(r)$ may be calculated as the smallest root of $H(a, r)$.

We may now solve the differential equation

$$\rho V(a, r) = q(a, r) + q(a, r)[V(a, r + 1) - V(a, r)] + \dot{V}(a, r) \tag{A3}$$

with boundary condition

$$V[t(r), r] = v[t(r)] \tag{A4}$$

recursively starting with $V(a, r^* - 1)$ and using $V(a, r^*) = q(a, r^*)\xi(a)$.

Risk Aversion Model

Here we assume worker preferences are represe
where u is strictly increasing and concave (and an
is also assumed), $w(a)$ is the real wage paid (and a
and ρ is the market interest rate. Let $v(a) = u(W$
$u(W)\xi(0)$. We assume that the cost of moving to
primary job sector is negligible. For the time bein
to the primary job is suboptimal. The downward
results implies that this is indeed the case using an
used under risk neutrality. It can be shown that an optim
schedule solves the following dynamic programming problem

$$V = \max \begin{cases} v(a), \\ V(0, a, r), \\ \left(\dfrac{1}{\rho}\right) \max\limits_{w, \Delta, e, \delta, \sigma} u(w) + q[V(x + \Delta, a, r + 1)e + v(1 - e) - V] \\ \qquad + V_x\delta + V_r + \sigma(v - V), \end{cases} \tag{A5}$$

subject to $w \geq 0$, $\sigma \geq 0$,

$$\rho x = w - q + qe(x + \Delta) - qx + \delta - \sigma x \tag{A6}$$

$$eV(x + \Delta, a, r + 1) = e \max\{V(0, a, r + 1), v\} \tag{A7}$$

with boundary conditions $V(x, A, r) = 0$ for all x, r, $X(0) = X(A) = 0$, $V(0, 0, 0) \geqslant v(0)$. In this problem, unless explicitly noted, all functions are evaluated at (x, a, r), (a, r), or a as appropriate. x is interpreted as the expected present value of the net cost of the employee to the firm; Δ is the amount by which x jumps if a unit of output is produced; e is the probability of retention immediately after a unit of output is produced; δ is the derivative of x with respect to a; and σ is the probability per unit time of not retaining the worker when no output is produced.

The first-order conditions can be written as

$$u'(w) = V_x, \tag{A8}$$

$$e\left[\left(1 + \frac{\theta}{q}\right) V_x(x + \Delta, a, r + 1) - V_x\right] = 0, \tag{A9}$$

$$\begin{gathered} e = 1 \text{ if and only if } V(x + \Delta, a, r + 1) - v - (x + \Delta)V_x \geqslant 0, \\ \text{and } 0 \text{ otherwise,} \end{gathered} \tag{A10}$$

$$V > v \text{ if and only if } V \geqslant v + xV_x. \tag{A11}$$

where $\theta \geqslant 0$ is a Lagrange multiplier for (A7). Note that (A8) and (A9) imply (since $\theta \geqslant 0$ and $\theta = 0$ whenever [A7] is slack)

$$u'[w^*(x + \Delta^*, a, r + 1)] = V_x(x + \Delta^*, a, r + 1) \leqslant V_x(x, a, r) = u'[w^*(x, a, r)]$$

with equality if (A7) is slack (asterisk indicates optimal value). Since $u'' < 0$, $w^*(x + \Delta^*, a, r + 1) \geqslant w^*(x, a, r)$ with equality if (A7) is slack. Thus, if r jumps at a, then, after optimally adjusting x by Δ^*, the new wage is the same as the old wage unless it must increase to keep the worker from moving.

It turns out that, if r does not jump at a, the wage will not change in a neighborhood of a. To see this, totally differentiate (A5) with respect to x assuming that $V > \max\{v, V(0, a, r)\}$. Using the envelope theorem and the first-order conditions, we obtain $\delta V_{xx} + V_{ax} = 0$. But using (A8), we have

$$u''(w^*)\frac{dw^*}{da} = V_{xx}\frac{dx^*}{da} + V_{xa} = V_{xx}\delta + V_{ax} = 0,$$

which implies that $dw^*/da = 0$ since $u'' < 0$. Thus wages are rigid except for upward jumps at points where r jumps and (A7) is binding.

References

Alchian, Armen A. "Private Property and the Relative Cost of Tenure." In *Economic Forces at Work*. Indianapolis: Liberty Press, 1977.

DeGroot, Morris H. *Optimal Statistical Decisions*. New York: McGraw-Hill, 1970.

Freeman, Smith. "Wage Trends as Performance Displays Productive Potential: A Model and Application to Academic Early Retirement." *Bell J. Econ.* 8 (Autumn 1977): 419–43.

Harris, Milton, and Holmstrom, Bengt. "A Theory of Wage Dynamics." *Rev. Econ. Studies* 49 (July 1982): 315–33.

Harris, Milton, and Weiss, Yoram. "Job Matching, Risk Aversion, and Tenure." Working Paper, Northwestern Univ., 1982.

Jovanovic, Boyan. "Job Matching and the Theory of Turnover." *J.P.E.* 87, no. 5, pt. 1 (October 1979): 972–90.

Jovanovic, Boyan, and Mincer, Jacob. "Labor Mobility and Wages." Working Paper, Columbia Univ., 1978.
Mortensen, Dale T. "Specific Capital and Labor Turnover." *Bell J. Econ.* 9 (Autumn 1978): 572–86.
Rothschild, Michael. "A Two-armed Bandit Theory of Market Pricing." *J Econ. Theory* 9 (October 1974): 185–202.
Viscusi, W. Kip. "Job Hazards and Worker Quit Rates: An Analysis of Adaptive Worker Behavior." *Internat. Econ. Rev.* 20 (February 1979): 29–58.
Weiss, Yoram. "Output Variability and Academic Labor Contracts." Working Paper no. 19-81, Foerder Inst. Econ. Res., Tel-Aviv Univ., 1981.

Book Reviews

La Théorie générale des surplus By MAURICE ALLAIS.

Paris: Institut de Sciences Mathématiques et Economiques Appliquées (ISMEA), Cahiers de l'ISMEA, ser. EM, no. 8, Economies et Sociétés, January–March 1981. Pp. 434. Fr 280 80

According to the French philosopher Edgar Morin, "great minds drill with unbending thought."[1] This aphorism is illustrated by Maurice Allais's book *La Théorie générale des surplus* (General theory of surpluses). The analysis relies on the concept of distributable surplus (*surplus distribuable*), which was introduced by the same author in 1943 in his *A la recherche d'une discipline économique*

Loosely speaking, distributable surplus of a given good for a set of consumption or production units, associated with a given modification of the economy starting from a given initial situation, can be defined as the amount of this good that could be made available after the change is made, under the conditions that all preference indices get values at least equal to those obtained in the initial situation, that resources used remain smaller than or equal to those of the initial situation, and that production made available by the group to the rest of the economy does not diminish The concept relies on the comparison of the initial state ξ_1 the state ξ_2 obtained after transformation and a state ξ_3 with the same utility levels as ξ_1 (*isohédone* in Allais's terminology), obtained from ξ_2 through changes of the allocation of the good in which surplus is expressed. The concept should be clarified and made more precise, in particular to distinguish between "distributable" and "distributed" surplus, "freed" (*libéré*) and "freeable" (*libérable*) surplus Deadweight loss of a given state is then nothing but the maximum distributable surplus obtainable from this state for the whole economy (this is also the definition in the 1943 book).

Allais's concepts of surplus and loss have a Hicksian flavor but are specifically designed to fit Pareto theory in a general equilibrium framework. Allais views his notion of surplus as a unifying concept in economic theory, whereas normative economics focuses attention particularly on the states in which distributable surplus is zero, positive economics consists in the description of the ways in which distributable surplus is "sought, realized, and shared." With this general remark in mind, it is convenient to consider separately three different aspects of Allais's new book First, it extends and unifies previous analytical work on a general equilibrium measurement of surplus. Second, it connects this technical effort with a synthetic reflection, placing the concept of surplus at the core of some of the modern developments of eco-

---

[1] *"Un grand esprit fore avec une idee fixe."*

[*Journal of Political Economy* 1984, vol 92 no 4]
© 1984 by The University of Chicago All rights reserved

nomics. Third, it provides a detailed and critical genealogy of the concept, over the century that followed Dupuit's seminal article. Each of these themes contains rich materials and leads to interesting achievements, and I will consider them successively.

I

Analytical achievements consist in the derivation of first- and second-order differentials of the distributable surplus from any initial state for any change of the economic variables compatible with the constraints It should be mentioned that the three different cases treated successively—measurement of surplus (1) when no surplus is generated from the production sector, or (2) when this latter condition is dropped, that is, for any modification of the economy (chap 2, pt. 2), or (3) when transformations are iso-utility (*isohédones*) transformations (chap. 3, pt 2)—are closely similar although different in terms of underlying economic processes In about one hundred and fifty dense pages (pts 1 and 2 and the beginning of pt 3), the author derives an impressive number of different formulas, among which a dozen should be singled out for their particular economic significance

The expressions obtained concern first- and second-order differentials but also finite approximations, they are valid for initial states outside of the vicinity of the first-best optimum, and hence they go further than the existing literature. Analytical tools are remarkably mastered and derivations are conducted with rigor and elegance I should, however, warn the reader that the matter is not always easy to grasp. Allais constantly uses second-order differentials. These tools, although adequate for the present problem, are somewhat unusual to economic theorists. To follow the derivations, a preliminary investment may be needed. One should also notice that the author systematically chooses to express results in terms of primal variables (quantities and derivatives of marginal rates of substitution and transformation with respect to certain directions in the quantity space, i.e., the Antonell
opposed to the Slutsky point of view for preferenc
ence to dual variables which are significant when m
market contexts. Comparison with the existing lit
mediate

II

The reflection on surplus appears in this book as a
ing different pieces of Allais's previous work As alre
of distributable surplus was presented in the 1943 book S
there had been already derived and many others were in filigree Both the Allais surplus and the related concept of loss were the starting points of further studies. The two articles published by Debreu in *Econometrica* (1951, 1954) on the subject are particularly notable The concept of welfare loss used by Debreu is grounded in Allais's surplus, but loss is measured with respect to a given composite commodity rather than to a noncomposite numeraire Debreu obtains second-order approximations of welfare loss in the neighborhood of a first-best optimum situation. The influence of Allais on the Boiteux study published in *Econometrica* (1951) has to be found in the techniques of second-order differentials; but his concept of welfare loss relies on the Hicksian compensating variation The different points of view coincide in specific cases and lead to identical second-order approximations of welfare loss only around the specific first-best optimum that is the social welfare optimum Allais's result at the end of chapter 1 (pt 3) on the invariance of second-order

approximations of losses with respect to the numeraire in a neighborhood of a first-best optimum (a point that has also been stressed by Diewert in a recent article [1981]) helps to clarify the relationships between different approaches.

Normative economics provides the logical thread that connects the different stages of Allais's reflection. The book reminds us of some early contributions on which evidence is brought in support of claims to priority, such as the introduction of the concept of social utility frontier introduced as *surface de rendement social maximum.* Also, classical questions of normative economics have answers appearing as the subproducts of a comprehensive analysis of distributable surplus. In particular, Allais presents a full characterization of first- and second-order conditions for first-best Pareto optima, which is valid both for the convex sector and for the nonconvex (respectively, *le secteur non différencié* and *le secteur différencié* in Allais's terminology). Although this material is not actually new, the restatement of the results, particularly those obtained by Allais himself for the conditions of coexistence of firms producing in a locally convex area and in a nonlocally convex area of their production sets, is certainly welcome.

The concept of distributable surplus provides a link between normative and positive Allaisian economics: "*Dans leur essence toutes les opérations économiques quelles qu'elles soient peuvent être considérées comme se ramenant à la recherche, à la réalisation et à la répartition du surplus distribuable*" (In essence any economic operation, whatever it may be, should be viewed as related to the search, realization and distribution of distributable surplus.) Allais's theory of *équilibre de marchés* (see Malinvaud 1972) contrasts with the Walrasian theory of *équilibre de marché* (singular). Although Allais does not seem to be totally explicit about whether, in the spirit of the Coase theorem, distributable surplus would actually be totally exhausted through economic activity, he emphasizes the role of distributable surplus as the analogue of "potential" in physics, or, if one prefers, as the natural "Liapounov function" of non-*tâtonnement* theory.

Allais's insistence on building a theory that is valid even at a distance from first-best Pareto-optimal states certainly reflects the needs of his approach to positive economics. It may also have been motivated by considerations germane to the analysis of constrained second-best situations. Although this latter point is not emphasized in the present book, preoccupations of this type are at least implicit in the work of the early fifties on surplus: Debreu's already quoted article refers to the "tax subsidy problem," while Boiteux's first attempts to give a general analytical solution to the pricing problems of public firms subject to a budget constraint seem to rely on the concept of surplus. Modern developments in second-best theory have given up the surplus approach. If such an approach were again undertaken, Allais's work would provide an adequate basis.

III

The last part of the book proposes a historical perspective of the subject. Unsurprisingly, Dupuit is praised as the precursor and Pareto, Barone, and Hotelling as the major followers; Samuelson and others are also evoked. Allais has strong feelings on the subject; he provides a critical assessment that is sometimes controversial but should certainly be of the highest interest to historians of economic analysis. Although I am not competent to judge, I cannot resist mentioning the discussion of the origin of the Edgeworth box, which according to Allais,[2] as well as to Jaffé previously, should be called the

---

[2] Allais, on the other hand, has the highest esteem for Edgeworth.

Pareto box. This anecdote also reveals the unsatisfactory state of the history of economic analysis.[3]

From the seminal work of Dupuit, the limits of which are partly in its partial equilibrium nature, a mature general equilibrium understanding of the surplus has emerged, and Allais's contribution was one of the final decisive steps. The description of the progress of the subject provided by Allais looks, on the whole, convincing. Besides some reservations on the understanding of Hotelling's contribution, the reader might have liked a more complete analysis of the postwar history (Debreu, Boiteux, Harberger) and of the more recent developments in the field. Consumer surplus (partial equilibrium) theory has been reassessed more satisfactorily thanks to the systematic use of duality theory. Bob Willig's (1976) article, "Consumer's Surplus without Apology," is often referred to for a restatement based on expenditure functions, a restatement to which many have contributed: Diamond, Deaton, Gorman, McFadden, and others. However, the corresponding concepts (equivalent variation or compensating variations) would lead, if employed in a general equilibrium context, to paradoxes which Boadway, Mishan, and more recently Schweizer and also Dierker and Lenninghaus (1983) have again identified.[4] The questions raised by Allais are unavoidable.

Allais has been an innovative pioneer as well as the influential master of some of the best economists educated in France in the postwar period (e.g., Boiteux, Debreu, Lesourne, Malinvaud). Rather than the language barrier, it is the displacement of the center of intellectual life on the American side of the Atlantic Ocean that is responsible for a decreased visibility of his work. This book on the general theory of surplus reminds the profession of a strong personality and of a message that had an unquestionable influence on the economic theory of the fifties and sixties. This message offering a yet unsurpassed understanding of surplus in a general equilibrium context remains up to date.

Centre d'Économie Quantitative et Comparative

References

Boiteux, Marcel. "Le Revenu distribuable et l

Econometrica 19, no. 2 (1951): 112–33.

Debreu, Gerard. "The Coefficient of Resource Uti

(July 1951): 273–92.

——— "A Classical Tax-Subsidy Problem." *Econometrica* 2

14–22.

Dierker, Egbert, and Lenninghaus, J. "Surplus Maximization and Pareto Optimality." Roneotyped. Bonn Univ., 1983.

Diewert, W. Erwin. "The Measurement of Deadweight Loss Revisited." *Econometrica* 49 (September 1981): 1225–44.

---

[3] In fact, Allais argues that contemporary history of economic analysis is still less known than past history. The improvement of the understanding of our recent history is a task of undoubted social interest. One may wonder whether it is conceivable that some institution organize research on the recent history of economic ideas. A precise reassessment of the genealogy of ideas would be of interest to the whole profession, not only to root rewards more firmly in facts.

[4] On this point the reader will find it useful to read "Aggregate Welfare Measures and Compensation Tests" by Schweizer (1983).

Malinvaud, Edmond *Lectures on Microeconomic Theory*. Translated by A. Silvey. Amsterdam: North-Holland, 1972.

Schweizer, Urs. "Aggregate Welfare Measures and Compensation Tests." Discussion Paper no 123. Bielefeld: Univ. Bielefeld, 1983.

Willig, Robert D. "Consumer's Surplus without Apology." *A.E.R.* 66 (September 1976): 589–97.

Macroeconomics after Keynes. A Reconsideration of the General Theory. By VICTORIA CHICK.

Cambridge, Mass.: M.I.T. Press, 1983. Pp. x + 374. $30.00 (cloth); $13.95 (paper).

Anyone looking at the title of this book would naturally expect to find a survey of the development of macroeconomics after the publication of the *General Theory* in 1936. At least so it was with me. But in fact it is not that kind of book at all. Victoria Chick's intention is much better described by the subtitle: "A Reconsideration of the *General Theory*." She has written an exegesis of and commentary on Keynes's great work, complete with a concordance showing which chapters of Keynes are discussed in which chapters of Chick. This is a *Guide to Keynes*. I am still puzzled by the title. Maybe it is meant in the sense that one could be said to have painted a "View of Toledo" after El Greco.

Chick says that, in her discontent with the standard textbooks and their version of "Keynesian economics," she was driven to try using the *General Theory* itself as the text in an intermediate macro course. That led her in turn to write this commentary, to guide her students back to the real Keynes. There is a minor industry of this sort that brings with it the obvious danger that the *General Theory* will become a kind of holy scripture; some "fundamentalists" already seem to ask of a proposition in macroeconomics not whether it is true or false but whether it is "truly Keynesian." I want to emphasize that Chick is not of this persuasion. She has a mind of her own; she is prepared to find the *General Theory* incomplete, or fuzzy, or erroneous in places. Above all, she makes the important point that Keynes was aiming to describe and understand the economy of the 1930s. He had no interest in "generality" as a modern mathematical economist would understand it. He meant that his theory was general only in the sense that it was not so special as the orthodoxy of the time. He insisted on the fruitlessness of trying to understand persistent mass unemployment with the aid of a theory that denies the possibility of persistent mass unemployment. (History repeats itself. The unemployment is much less nowadays but the fruitlessness is just the same.) This analytical specificity has a cost: when the structure of the economy changes because its institutions or its experience have changed, some parts of a once valid theory are likely to become obsolete. Chick does not hesitate to apply this maxim to Keynes.

There is a bit of the fundamentalist in her, however. She is one of those who think that modern "Keynesian economics"—especially that old devil IS-LM—has somehow betrayed the *General Theory*. Now there is nothing wrong with that attitude in principle, although I do think it gives inadequate weight to the extent to which any substantial body of science is a collective product. This is as true of Keynesian economics—without the quotation marks—as it is of Newtonian physics or the Darwinian theory of evolution. It could easily be the case that a pioneering work like the *General Theory* should contain important half-formulated insights that do not get incorporated into the early formalizations and thus tend to sink from view. In fact I have a certain amount of sympathy for some of Chick's arguments in this vein. Let me give two important examples. She says that modern Keynesianism has tended to submerge the distinction between investment spending and consumption spending and their different roles in economic fluctuations. She claims that this is certainly a distortion of Keynes's intentions, and a mistake besides. I think there is much truth in that. The change came about in the course of the (somewhat excessive?) fuss over "microfoundations," probably because of the mathematical economist's tendency to speak in terms of abstract agents and commodities distinguished only by subscript. This criticism is not true of all modern Keynesianism, but I would agree that reemphasis is called for.

The second example is the treatment of uncertainty. There is little doubt that Keynes himself was usually thinking of "Knightian uncertainty" and not probabilistic risk when he talked about the intertemporal decisions of firms. He would have dismissed the modern penchant for describing the world in terms of (stationary!) stochastic equilibrium as simply irrelevant, incapable of describing the sort of thing that actually happens. Chick is right to insist on that. Of course Keynes might have been wrong in his judgment, but that needs to be argued back and forth. There is a two-edged blade here. Those who say that the probability calculus is all we've got, so we might as well use it, have a point, but they forfeit their right to it when, five steps later, they have completely forgotten that they have built on a major unargued premise. Nor can they really rely on the Ramsey-Savage axiomatization of subjective probability. I can be forced, kicking and screaming that I really do not know, to reveal a subjective probability of one-half for event *E*, but you cannot guarantee that I will then go home and make investment decisions as if *E* were determined in fact by the toss of a fair coin. I may not flinch [illegible] have violated one of the axioms. On the other side, [illegible] about "leaps into the unknown" instead of probabili [illegible] a lot of effort in finding a reasonably coherent way [illegible] tainty. Otherwise a lot of territory has to be aban [illegible] retreat into Delphic utterance and answer a real qu [illegible] question, as I fear Chick occasionally does. (She [illegible] vaguely hostile to theories couched explicitly as su [illegible] method she describes as being suitable only to the a [illegible] presumably not to the analysis of production. I wonder if s [illegible] herself what a theory determining the values of two or more variables [illegible] possibly look like, under the skin, if not simultaneous equations, perhaps thinly or thickly disguised.)

Keynes claimed that the industrial capitalist economies of his time could not be relied on to find their way to full-employment equilibrium and stay there under the inevitable shocks, mostly real shocks. He located the source of this failure in (at least) two places. One is the labor market; something must be going wrong there, or unemployment would not be the characteristic pathology of recession. The other is the capital market, something about the rate of interest failing to match saving and investment at full employment. Chick very sensibly pays attention to both themes and has many useful things to say.

On the labor market she is fairly conventional, arguing in the usual way that institutional arrangements and the inevitable characteristics of a decentralized many-commodity economy make it impossible for workers to initiate bargains over the real wage. Nor may employers wish to bring about reductions in the real wage, even in recession, since they have to worry about the reactions of their already employed workers. Chick emphasizes a point often overlooked, that employed and unemployed workers have divergent interests at such times, and this too works against real-wage flexibility. "It can be seen that there is little justification for expecting wages to fall in periods of unemployment, at least not until the unemployment and the low profits associated with the low level of demand for output have persisted for quite a time." And even then, falling wages may not do much good if nominal demand for output falls with them. She does make a bit of heavy weather over the difference between "rigid" or "given" and "sticky" wages. "The wage is neither given nor endogenously determined." What then? Well, perhaps it is "determined by history." But that is what one means by "given." The important thing is to agree that wage rates do not move flexibly to clear labor markets quickly, for all the nitty-gritty reasons she mentions and others besides.

On saving and investment, her starting point is Keynes's observation that an act of saving is a decision to forgo consumption of more or less specific things now; even if it entails an implicit decision to buy something later, both the time and the object of expenditure are vague and uncertain. (This is where the distinction between risk and uncertainty comes in. In contemplating the investments that are needed to offset saving, firms may not act as if they are buying a well-defined lottery ticket over possible times and objects of consumer spending.)

I have already admired Chick's faithfulness to the idea of investment spending as the driving force in economic fluctuations. Presumably she would have no difficulty in adding some consumer durables, for example, to this class of active expenditures; this would be an instance of the need to modify theory to fit the facts of changed economic structure. Nevertheless, I found her treatment of saving and investment naggingly insufficient, in many small ways rather than one big way I have space to mention only a few. She takes an oddly literal-minded view of some standard moves in dealing with investment. For example, she dislikes the conventional idea that ex ante saving decisions are always carried out so that the full burden of accommodating differences between intended saving and investment falls on unintended inventory accumulation. Her objection is that this makes investment seem passive and not the causal agent. It is not hard to allow realized saving and investment to fall anywhere between the ex ante magnitudes if there is any good reason to do so, but the reason she gives seems to turn on unimportant appearances, not substance.

Again, she criticizes the multiplier-accelerator model as being too simplistic to represent the full expectational-financial character of Keynes's thinking about investment. But the model was not proposed as a complete theory of investment, only as an illustration of the important fact that stock-flow interactions contain many dynamical possibilities, including cyclical ones. To my mind, this excessive literalness also interferes with her discussion of the multiplier. She somehow manages to make a difference of principle between the "step multiplier" and the "blip multiplier"—those are my self-explanatory shorthand names—where I see none; and I am dubious about her discussion of the monetary accompaniment of multiplier processes.

While I am being critical, let me give one last example of what I have called "literal mindedness." Chick is deeply hostile to the standard Keynesian model as descended from Modigliani, the version that appends to the IS-LM equations the simplest possible supply side: short-run production function, a price-equals-marginal-cost condition, and a labor supply equation (or a given nominal wage). She seems to believe that just because it may be algebraically convenient to solve these equations by starting with the labor market, the model therefore attributes "causal priority" to the labor market. Now I think I actually agree with her about many of the specific deficiencies of the model, but I do not see where causal priority comes into it. No wonder she has such a hard time with simultaneous equations.

This book does not represent my favorite way to do macroeconomics, as the reader will have gathered. I think formal model building is both a good way to exercise the imagination and an important discipline on the imagination. I wish Chick's approach were more technical. But I hope the reader has also gathered that I found it a provocative and interesting piece of work. Chick cares about many of the right things, pursues them tenaciously, and is not bemused by fashionable fluff. Her readers will be led, and frequently helped, to think about serious issues of macroeconomic theory.

ROBERT M. SOLOW

Massachusetts Institute of Technology

How We Live: An Economic Perspective on Americans from Birth to Death. By VICTOR R. FUCHS.
Cambridge, Mass.: Harvard University Press, 1983. Pp. 293. $17.50.

Friedrich Engels and Joseph Schumpeter both forecast a decline of the conjugal family, and both anticipated significant consequences from that decline. In his latest book, a synthesis of a vast array of empirical studies documenting how Americans live, work, procreate, and die, Victor Fuchs p
tial evidence that the family is declining, explores th
this reviewer that the consequences are significant

The subject Fuchs has chosen is important, and
book can be recommended for scholars, students, p
friends or relatives who ask "What do economists do
and nontechnical language what economists have be
to describe and understand modern family life.

This book provides an economic perspective on
death. It is about how we choose to live, the relationsh
made by us and by others, and the constraints we face. The book is organized in six chapters, each dedicated to a different phase of the life cycle, plus a concluding chapter on policy. Evidence from scores of economic, demographic, and medical studies is examined in the context of several major themes: changes in external circumstances, the interrelatedness of choices, the ways in which demography transforms circumstances facing individuals, and the significance of the ability to delay gratification.

The studies cited in this volume vary in quality, but the overall standards and cumulative development of evidence are impressive. Rather than attempt to enumerate the wide range of topics and studies covered, I will illustrate Fuchs's approach with some examples of his analysis of parent-child relationships.

In the chapter on birth and infancy, one issue examined is infant mortality. The infant mortality rate in 1915 was 100 per 1,000 live births. Since then the rate has been decreasing at more than 3 percent per annum, reaching 12 deaths per 1,000 live births in 1981. Although overall rates have been falling, the nonwhite rates have remained 60–90 percent higher than the white rates since 1940, with no clear trend.

In examining this differential, Fuchs points out that age, education, and marital status of the mother and the timing and amount of prenatal care differ by race, but the differences are not large enough to explain the difference in distribution of birth weights by race, which is the most important "explanation" for differences in infant mortality by race. Mortality rises as birth weight falls: babies weighing under 2,500 grams (5.5 pounds) account for only 7 percent of all births but for half of all neonatal deaths. Thirteen percent of all black babies weigh less than 2,500 grams compared to only 6 percent of white babies. About two-thirds of the race differential in infant mortality is eliminated by standardizing for differences in birth weight.

Delay in receiving prenatal care is associated with lower birth weight, but the impact is relatively small. A delay of 6 months in seeking prenatal care on average reduces birth weight by 1.5 ounces. This contrasts with the impact of cigarette smoking. One pack per day smoked by the mother during pregnancy lowers average birth weight by 1 pound, approximately doubling the probability of having a high-risk, low-weight baby. The importance of this startling relationship is emphasized by the additional finding that among babies born weighing 751–1,000 grams, about one out of every four survivors will have at least one major handicap, such as very low IQ, hearing loss, or spastic disorder (Fuchs does not propose policies to reduce smoking, but where is there a stronger case for taxes to reduce spillover effects?) Smoking does not help explain the lower birth weights of black babies since, in the sample examined, black women smoked less than whites.

Alternative explanations for birth weight differences by race are considered. Fuchs notes that 9 percent of Puerto Rican babies weigh less than 2,500 grams at birth, while only 5 percent of Mexican babies are in this high-risk category. In Texas, the Spanish-surname population, which is mostly of Mexican origin, enjoys the same low infant mortality as the white Anglo population. In New York, infant mortality among Puerto Rican Hispanics is much higher than among non-Hispanic whites. Differences in income, education, access to medical care, language, or religion cannot be the explanation. Fuchs argues that one major difference is in family stability: women of Puerto Rican origin are much less likely to have grown up in a stable family and are less likely to be married at the time they give birth. Fuchs concludes, "Perhaps fetal development depends upon strong social support from family and community" (p. 37).

The conditions and circumstances of birth are relevant to a range of issues, including the distribution of income both for children and for mothers. In the chapter on adolescence, Fuchs examines the increased demand for children by unmarried mothers. One study of pregnant teenage girls found those with higher grades and those enrolled in school were more likely to abort. Catholics and non-Catholics, other than Mexican Americans, did not differ in their choices. However, when age, school enrollment status, grades, religion, and ethnicity are held constant, the girls eligible for public aid were significantly more likely to deliver out of wedlock. Incentives matter here as elsewhere. Fuchs argues that sensible public policy should be directed toward providing these girls more attractive educational and employment alternatives.

Changes in family life among young adults are also documented. Currently about half of all weddings in the United States involve at least one person who has been married before. In 1950 nine out of 10 married women with children under 6 years of age were not working for pay. Half were so employed by 1980. The reasons for this rapid growth are discussed. Fuchs argues that the feminist movement, rising divorce rates, and government affirmative action programs cannot be the causes, since they did not gain force until well into the 1960s, whereas the rapid rise in participation of married women began in the 1950s. Fuchs's own explanation emphasizes the growth of real wages and expansion of the service sector.

Concerning male-female wage differentials, Fuchs argues that "parenthood" is the most serious barrier to occupational equality. In one of the more depressing statements of the book, Fuchs argues that there is "one large obstacle remaining in the path of economic equality for women: their demand for and concern about children. If, as current behavior suggests, this demand and concern are not as great among men as among women, two outcomes are possible. On the one hand, concern for children may prevent some women from committing themselves as wholeheartedly as men to paid careers, and an earnings gap will remain. On the other hand, full commitment to careers by women may result in further decline in fertility and less parental attention to children. . . . To the extent that women become less interested in bearing children or less concerned about their welfare after birth, the future of the family and of society looks bleak because there is little evidence that men are prepared to take up the slack" (pp. 143, 158).

Discussing the reasons for these changes in family life, Fuchs considers the growth of government programs that have altered incentives and facilitated changes in family structure. Fuchs's own theory of why government programs have expanded into traditional family responsibilities is that "the driving force behind most social legislation in the United States and other developed countries is the pressure for a more egalitarian society. [illegible]
modern life compel a more equal sharing of [illegible]
power. . . . Modern governments use egalitarian [illegible]
induce mass participation and cooperation in the [illegible]

The argument that government programs [illegible]
groups is unexceptionable. The implication that the [illegible]
are primarily motivated by a desire to reduce [illegible]
vincing. Most of what is redistributed does not go to [illegible]
affected by who votes and how interest groups are [illegible]
nizes, if children voted, our policies would differ from the [illegible]
egalitarian theory remains to be established as a satisfactory theory of the growth of government; and the issue more specific to this book remains to be resolved: How much of the expansion in government welfare programs is in response to the declining family and how much is the cause?

One particularly interesting issue raised by Fuchs concerns the role of differential discount rates across individuals in "explaining" such behavior as smoking, investment in education, and the positive correlation between mothers' education and investment in children. One study found that "the ability to delay gratification during childhood is positively related to the presence of a father in the home as well as to intelligence and achievement orientation." Given the potential significance of this variable, the amount of available evidence appears modest. What do we know about systematic influences on individual discount rates? Very few economists have directed their attention to this issue, and the case Fuchs makes is sufficiently strong to suggest underinvestment in this area.

Fuchs's analysis and evidence raise the standards of debate. Sensible discussion of a wide range of issues, from the distribution of income to changes in fertility and employment patterns and the efficacy of government social insurance programs, must confront the evidence and arguments presented here.

In our specialized society, the life cycle now moves from nursery school to nursing home. The implications of such changes in family structure are not obviously coincident with those suggested by either Schumpeter or Engels but are nonetheless profound. Few issues that concern economists are unaffected by the patterns Fuchs documents

LEE BENHAM

Washington University

Violations of the Gold Points, 1890–1908

Truman A. Clark

University of Southern California

Evidence is presented that the gold standard system was inefficient in two ways. (1) Gold point violations sometimes persisted for long periods indicating that profit opportunities were not always eliminated quickly. (2) Following gold point violations, gold occasionally flowed in unprofitable directions. It is argued that government interference in the workings of the system was a source of these inefficiencies. This suggests that the gold standard did not completely restrict the discretionary power of monetary authorities. If present-day gold standard advocates seek to eliminate discretionary monetary policy, revival of the traditional gold standard is probably insufficient to attain their goal

The current debate concerning the desirab
form of a gold standard has stimulated renew
standard system of the late nineteenth and ea
Under that system, market exchange rates
official exchange rates within variable bands
of shipping gold. The limits of those band
profitable to ship gold between countries, were
points." Private gold shipments in response to gold
served to limit the discretionary power of monetary authorities over national money supplies. The desire to establish similar limits on the discretionary power of today's central banks is a main motivation for proposals to restore a gold standard.

This study is concerned with the efficiency of the gold standard

---

This paper has been improved by the helpful comments of Fischer Black, Douglas Joines, Merton Miller, Marc Reinganum, Mark Weinstein, and a conscientious referee. The research was supported by a grant from the School of Business Administration, University of Southern California.

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0011$01.50

system. It focuses on the system's reaction to violations of the gold points between the United States and Great Britain during the years 1890–1908. After the occasions on which gold point violations occurred are estimated, two aspects of the efficiency of the gold standard system are assessed: the speed with which gold point violations were eliminated and the direction of gold flows in response to profit opportunities. It is found that gold point violations sometimes persisted for prolonged periods and that gold often flowed in apparently unprofitable directions. It is argued that a likely source of many of these inefficiencies was government interference in the workings of the gold standard system. While the gold standard set long-run limits on their actions, the Bank of England and the U.S. Treasury possessed some discretionary powers. Those powers occasionally were used to promote gold imports, to block gold exports, and to delay the workings of the gold standard system.

The paper is organized into six sections. In Section I, the conditions for the profitability of gold shipments between the United States and Great Britain are derived. Estimates of gold shipment profits occurring between 1890 and 1908 are reported in Section II. The sensitivity of the estimated profits on gold shipments to several types of errors is considered in Section III. Section IV is concerned with the efficiency of the gold standard system. Section V discusses government intervention in the gold market. The conclusions of this study are presented in Section VI.

I. The Profitability of Gold Shipments

Under the gold standard, the profitability of shipping gold between the United States and Great Britain depended on market exchange rates, the official exchange rate, and the cost of shipping gold. The official exchange rate (X/Y), or "mint parity," was determined by the dollar price (X) that the U.S. Treasury was required to pay or receive per unit of gold times the quantity of gold ($1/Y$) for which the Bank of England would pay or receive one pound sterling. The spot (S_t) and forward (F_t) market exchange rates (expressed in dollars per pound) fluctuated about mint parity (X/Y) within bounds determined by the cost of shipping gold. There were two components of that cost: the direct costs (which included charges for transportation, safekeeping, insurance, etc.) and the opportunity cost of the funds committed to the gold shipment operation. Forgone interest in the United States ($i_{US,t}$) or in Britain ($i_{UK,t}$) constituted the opportunity cost component. Gold shipping operations also required conversions between dollars and pounds and, in some cases, securities transactions. Such transactions generated commission expenses that included the necessity of trading across dealers' bid-ask spreads. However, because of the lack of estimates of their magnitudes, commission expenses are ignored in the following analysis.

For an American citizen importing gold into the United States from Britain, the first step (taken at time t) was the conversion of dollars into pounds at the spot exchange rate yielding $1/S_t$ pounds per dollar. Next these pounds were converted into gold at the Bank of England at the rate of $1/S_tY$ ounces of gold per initial dollar. The gold was then shipped to the United States. On its arrival, it could be converted back into dollars at the U.S. Treasury yielding X/S_tY per dollar originally invested. During the time it took to complete the entire operation (k days), an investment in American securities would have yielded $1 + i_{US,t}$ per dollar (where $i_{US,t}$ denotes the interest rate at time t of an American security maturing at $t + k$).[1] Expressing the direct shipping costs (s) as a percentage of the initial dollar investment, the importation of gold would be profitable if the following inequality were satisfied:

$$\frac{X}{S_tY} > 1 + i_{US,t} + s \qquad (2)$$

The profitability of exporting gold from the United States to Britain involved similar considerations, but the appropriate exchange rate differed because the conversion of currencies occurred after the gold arrived at its destination. In order to reduce the danger of ad

---

[1] In the absence of transactions costs and market imp conditions" would have to hold for there to be no n opportunities. To express these conditions in terms of rates, let $i_{US,t}$ denote the rate of interest at time t on Ame k, and let $i_{UK,t}$ denote the rate of interest at time t on Britis that also mature at $t + k$. Also, let S_t denote the spot (doll and let F_t indicate the forward (dollar/pound) exchange rat sion to occur at $t + k$. Equation (1) expresses the relation betwe interest rates that must hold in order that the expected return on American and British securities may be equal

$$\frac{F_t - S_t}{S_t} = \frac{i_{US,t} - i_{UK,t}}{1 + i_{UK,t}} \qquad (1)$$

For a derivation of the interest parity conditions, see Cohen (1969, pp. 64–67). Assuming that the interest parity conditions as expressed in eq. (1) held, an American engaged in gold shipping had to consider only the interest rate on American securities. If the interest parity conditions were violated, an arbitrage opportunity in international securities trading would arise, but the securities transactions involved would not necessitate any gold shipments. Thus, a potential profit on the difference between American and British interest rates was a separate matter that did not affect the profitability of gold shipments directly. Only interest rates in the United States directly influenced American decisions to export or import gold, and, similarly, only British interest rates entered into the profit calculations of Britons contemplating gold shipments.

verse changes in the spot exchange rate during the time the gold was in transit, a forward exchange contract was necessary.[2]

The first step in a gold export operation (begun at time t) was the purchase of gold from the U.S. Treasury at the rate of $1/X$ ounces per dollar. On the same date, a forward contract was obtained to convert Y/X pounds (per initial dollar) into dollars at the rate of F_t dollars per pound on the date the gold was expected to arrive in Britain. The gold was shipped to Britain, and, on arrival, it could be presented at the Bank of England for conversion into Y/X pounds (per initial dollar). The pounds would then be delivered against the forward contract, yielding F_tY/X dollars per dollar originally invested. The total cost of the operation was the sum of the direct costs (s)[3] and the interest cost ($1 + i_{US,t}$). Exporting gold was profitable if the following inequality held:

$$\frac{F_tY}{X} > 1 + i_{US,t} + s. \tag{3}$$

By similar reasoning, the conditions for the profitability of gold shipments undertaken by British subjects can be derived. Gold imports from the United States were profitable whenever the following inequality was satisfied:

$$\frac{S_tY}{X} > 1 + i_{UK,t} + s. \tag{4}$$

Gold exports to the United States were profitable if

$$\frac{X}{F_tY} > 1 + i_{UK,t} + s.^{4} \tag{5}$$

---

[2] A gold exporter could choose not to close his position with a forward contract and take his chances on converting currencies at the spot rate prevailing when the gold arrived. By doing so, he would be speculating on exchange rate changes during the shipping period. Such exchange rate speculation could be undertaken directly at any time without the need to transport gold. Thus, the gain or loss on exchange rate speculation was a separate consideration, and the calculation of the profit attributable directly to a gold export operation properly involved the exchange rate of a forward contract used to close the position. However, as I will discuss below, even the use of forward contracts could not completely shield a gold exporter from exchange rate uncertainty because of the variability of transatlantic shipping times.

[3] For simplicity, the direct costs of gold shipments are assumed to be constant over time. In fact, these costs were declining secularly (see Myers 1931, p. 342). It is also assumed that the direct costs of shipping gold are the same in each direction and that they are independent of the number of days of shipping time.

[4] Examining (2) through (5), one can see that the cost of shipping gold set limits on the extent to which the spot and forward market exchange rates could differ from mint parity. These limits were known as the "gold points," and they can be derived from (2) through (5). For example, define $IM_{US,t}$ as follows:

$$IM_{US,t} = \frac{X}{Y}\left(\frac{1}{1 + i_{US,t} + s}\right). \tag{6}$$

Referring to (2) and (4), gold importing (into either the United States or Britain) involved no exchange rate uncertainty since currency conversions took place at the spot rate prevailing at the start of the operation. The only source of uncertainty was the time that would elapse while the gold was in transit. Undoubtedly, much of this uncertainty was attributable to the weather. If everything else remained the same, the profit on gold imports was greater as the number of days used in transportation declined since the interest cost was reduced. However, an unexpected storm that delayed transatlantic shipping by even a day could convert a profit into a loss.

The variability of shipping time also introduced the same element of uncertainty into gold export operations with regard to the interest cost. In addition, the variability of transportation time exposed gold exporters to a degree of exchange rate uncertainty. In a closed gold export operation, a forward contract setting the exchange rate for a currency conversion to take place on a specific future date was acquired. If the gold shipment arrived before the forward contract expired, the gold exporter could either (*a*) sell the contract before its expiration (which could produce a gain or a loss) and convert currencies at the spot rate or (*b*) invest the foreign currency received in exchange for gold in securities and then convert currencies when the forward contract came due. If the gold did not arrive by the time the forward contract came due, the exporter would have to liquidate the contract (which could entail a gain or a loss) and then either (*a*) enter into another forward contract to reduce further exposure to exchange rate fluctuations or (*b*) maintain an open
to convert currencies at whatever the spot
be when the gold finally arrived. Thus, due
atlantic shipping time, even the use of
fully eliminate exchange rate uncertainty
tions. Despite these elements of uncertainty
pants had strong incentives to ship gold
change rates violated one of the gold points

Then, rearranging the terms in (2) and substituting from the definition of IM_{US}, we see that it was profitable for an American citizen to import gold from Britain if S IM_{US}. Thus, IM_{US} represents the critical value of the spot exchange rate below which it became profitable to import gold into the United States, and IM_{US} was referred to as the American "gold import point." By similar substitutions, expressions for the American gold export point and the British gold import and export points can be derived. As indicated by (2) through (5), the gold points were variables rather than constants. However, it was apparently common around the turn of the century to quote fixed values of the exchange rate as *the* gold points. For example, Morgenstern (1959, pp. 176–80) observed that the *Economist* reported unchanging estimates of both the British gold export point (4.827) and the British gold import point (4.890) every week during the period from September 22, 1877, to August 19, 1916. For a review of contemporary estimates of the gold points, see Morgenstern (1959, pp. 177–81).

II. Estimates of the Profits on Gold Shipments, 1890–1908

The frequency of gold point violations during the gold standard era has been examined before. As part of his comprehensive study of international financial markets, Morgenstern (1959, pp. 245–46, table 53) identified 45 months between January 1880 and July 1914 when violations of the British-American gold points apparently occurred. However, Morgenstern's findings were questioned by Borts (1960) for several reasons.

First, Morgenstern used constant values of $4.845 for the British gold export point and $4.890 for the British gold import point throughout the 1880–1914 period. (He derived these figures by taking the medians of several gold point estimates published by others.) As Borts (1960, pp. 225–26) pointed out, and as my equations (2) through (5) indicate, the gold points were variables and not constants. Moreover, the profitability of gold exports depended on forward exchange rates (or alternative covering transactions) and not spot exchange rates as used by Morgenstern.

Second, Borts (1960, p. 225) suggested that the gold points used by Morgenstern were biased because he omitted or underestimated important components of the costs of shipping gold. Among other things, Borts indicated that commission expenses were probably ignored and that an unrealistically short shipping time of 8 days was probably assumed. As a result, Borts maintained that the true gold points usually lay outside the range of $4.845–$4.890 used by Morgenstern. If Borts is correct, some (or perhaps all) of the gold point violations identified by Morgenstern may have been spurious.

However, Borts did not comment on one aspect of Morgenstern's data that would tend to offset the use of biased gold point estimates. According to Morgenstern (1959, p. 277), the exchange rate data he used were monthly averages of weekly observations. Since averaging will smooth out extreme weekly values within a month, the use of monthly average exchange rate data is likely to lead to an underestimate of the frequency of gold point violations.

My estimation procedures and data should avoid a number of the ambiguities of Morgenstern's estimates of gold point violations. Specifically, the profitability of shipping gold is estimated at monthly or weekly intervals with current market data. Gold export profits are estimated by means of a sequence of transactions designed to reduce the danger posed by adverse exchange rate changes. Profits are estimated over a range of shipping times from 7 through 14 days. Finally, weekly exchange rate observations are used instead of monthly averages. The only remaining deficiency of Morgenstern's estimates not corrected in this study is the omission of commission expenses from the profitability calculations.

The data that are available permit the estimation of the profitability of gold shipments monthly from 1890 through 1898 and weekly from 1899 through 1908. The spot exchange rate data used were the London rate of exchange on New York on the first Friday of each month from 1890 through 1898 and the London rate of (cable transfer) exchange on New York at the end of each week from 1899 through 1908. On each date, only a single exchange rate figure was available, so the cost of trading across the dealers' bid-ask spread could not be incorporated into the profit estimates. Whether the figures collected were discrete quotations, weekly averages of daily quotations, or some other transformation of actual quotations was not explained by the source.[5]

During the period 1890–1913, it reportedly took between 7 and 14 days for gold to be shipped from New York to London; the average was 10 days (Goodhart 1969, p. 110).[6] Estimates of the interest cost were derived from the weekly average call money rates in New York and London.[7] It was assumed that the funds employed in gold shipping could have been invested at the weekly average call money rates during the time the gold was in transit.

For gold shipped on day t and arriving k days later (at $t + k$), the interest cost was computed as follows:

$$1 + i_{j,t} = (1 + a_{j,t+7})^{7/365}(1 + a_{j,t+14})^{(k-7)/365}$$

In equation (7), the subscript j identifies the
and $a_{j,t+7}$ and $a_{j,t+14}$ represent the weekly
(expressed on a per annum basis) in country
following the shipping date. The interest co
times were estimated by letting k take on val

In 1906, direct shipping costs were repor
cent (Myers 1931, p. 343). With the passage
1873, the United States established the official exchange
$4.8665 per pound sterling. These figures can be substituted into (2)

---

[5] Both of these data sets were taken from U.S. Congress (1910): the monthly data were reported in table 3, pp. 67–69, and the weekly data appeared in table 4, pp. 70–74. The original source for both data sets was the *Economist*.

[6] As early as 1893, it was apparently common in New York City to quote prices for gold to be delivered in 10 days (see Sprague 1910, p. 194).

[7] The U.S. weekly average call money rate was taken from Andrew (1910, pp. 119–37, table 29). The British weekly average call money rate came from U.S. Congress (1910, pp. 44–62, table 18).

and (4) to produce the following representations of the gold import profitability conditions:

$$\frac{4.8665}{S_t} - (1 + i_{US,t}) - .003125 > 0; \tag{2'}$$

$$\frac{S_t}{4.8665} - (1 + i_{UK,t}) - .003125 > 0. \tag{4'}$$

The lack of market forward exchange rate data makes it more difficult to estimate the profitability of gold exports, but the problem can be overcome by tracing the actions a gold exporter could take in lieu of forward contracting to reduce his exposure to exchange rate uncertainty. For example, an American gold exporter could first borrow pounds in London on the shipping date (t) at the rate $i_{UK,t}$. The amount in pounds borrowed at t would equal the present value of the amount in pounds to be received when the gold arrived in London. The discount rate used to obtain this present value would be $i_{UK,t}$. The pounds borrowed at t could next be converted into dollars at the spot exchange rate (S_t) and invested in American securities at the rate $i_{US,t}$. When the gold arrived in London, it would be converted into pounds, which would then be used to repay the loan. The American securities held could then be liquidated to realize the profit on the gold export operation in dollars. These transactions reduce the American gold exporter's exposure to exchange rate uncertainty because the currency conversion takes place at the known spot rate (S_t) on the shipping date. However, as with an explicit forward contract, the gold exporter would still be exposed to a degree of exchange rate uncertainty because of the variability of the shipping time.

By a similar combination of transactions involving borrowing dollars in the United States at $i_{US,t}$, converting those dollars into sterling at the spot exchange rate (S_t), and then investing the pounds in British securities at the rate $i_{UK,t}$, a British gold exporter could also reduce his exposure to the danger of adverse exchange rate changes. The implicit forward rate (F_t^*) in these transactions is expressed in equation (8):

$$F_t^* = \frac{S_t(1 + i_{US,t})}{1 + i_{UK,t}}. \tag{8}$$

The virtue of the implicit forward rate is that it can be computed from the available interest rate and spot exchange rate data.[8]

---

[8] The set of transactions assumed in deriving the implicit forward rate is referred to as a "money market hedge" in the international finance literature. I am indebted to Alan C. Shapiro for bringing this to my attention. It is interesting to note that the implicit forward rate (F_t^*) in (8) will equal the market forward exchange rate (F_t)

When the implicit forward rate is substituted into (3) and (5) along with the official exchange rate and the estimated direct shipping costs, the gold export profitability conditions become

$$\frac{S_t(1 + i_{US,t})}{4.8665(1 + i_{UK,t})} - (1 + i_{US,t}) - .003125 > 0 \qquad (3')$$

and

$$\frac{4.8665(1 + i_{UK,t})}{S_t(1 + i_{US,t})} - (1 + i_{UK,t}) - .003125 > 0. \qquad (5')$$

By means of the interest rate and spot exchange rate data sets described above, (2′), (3′), (4′), and (5′) were used to estimate the profits on gold shipment operations.

In tables 1 and 2, the dates when the estimated rates of return on gold shipments between the United States and Britain were positive are reported.[9] The figures were computed on the assumption

---

whenever the interest parity conditions of eq. (1) hold. Another alternative to forward contracting to cover gold export operations was mentioned by Borts (1960, p. 226). It called for banks in the gold exporting country to sell (for domestic currency) sight drafts drawn in the foreign currency on banks in the importing country. The drafts sold would be redeemed in foreign currency when the gold export operation was completed. If data on sight drafts can be found, it would be interesting to estimate the profitability of gold exports covered in this manner.

[9] Five dates on which profit opportunities were initially indicated are not reported in tables 1 and 2. It appeared to be profitable to ship gold from the United States to Great Britain on November 9, 16, and 23, 1907. However, these [illegible] 1907 when currency premiums arose in the United S[illegible] miums are considered, American gold exports in Nov[illegible] in losses. The effects of currency premiums on the p[illegible] discussed in App. A below. Two other dates were [illegible] exchange rate quotes. On all dates when profit oppo[illegible] change rate quotes obtained from U.S. Congress (191[illegible] table 4) were checked for consistency with exchange [illegible] first was Business Cycle Series 14.105, "Average Mon[illegible] on New York," compiled by the N.B.E.R. The month[illegible] computed from figures appearing in the *Economist*. It should be [illegible] was also the original source for U.S. Congress (1910, tables 3 and 4). The second source used in checking exchange rate quotes was Andrew (1910, pp. 189–94, table 13, pp. 195–207, table 14). Andrew's data were daily observations of the sight exchange rate for bankers' bills obtained from the *Commercial and Financial Chronicle*. For February 1, 1890, U.S. Congress (1910, table 3) quotes the exchange rate at 4.835. The use of this quote indicates that it was profitable to ship gold from Britain to the United States. The monthly average figures from N.B.E.R. series 14.105 are 4.8631 for January 1890 and 4.8688 for February 1890. For the week January 27–February 1, 1890, Andrew (1910, table 13) reports daily quotes falling in the range 4.88–4.885. Rather than its being profitable to ship gold to the United States from Great Britain on February 1, 1890, Andrew's quotes suggest it was almost profitable to ship gold in the opposite direction. For October 3, 1891, U.S. Congress (1910, table 3) quotes the exchange rate at 4.8875. The use of this figure indicates that it was profitable to ship gold from the United States to Britain. The monthly average N.B.E.R. figures are 4.8431 for September 1891 and

that it took 10 days to complete each gold arbitrage operation, and the profit rates are expressed as percentages.[10]

Table 1 presents the estimated profit rate derived from the monthly data for the period 1890–98. During those 108 months, there were 40 occasions when it was profitable to ship gold from the United States to Britain. The estimated profit rates on British gold imports ranged between 0.52 percent in August 1895 and 0.002 percent in May 1897. On each date on which it was profitable to import gold into Britain, it was also profitable to export gold from the United States, and there is a very close correspondence between the estimated profit rates on British gold imports and American gold exports on each occasion.[11] There were only seven dates on which it was profitable to ship gold from Britain to the United States. The estimated profit rates on American gold imports were as high as 0.10 percent in April 1898 and as low as 0.0004 percent in October 1897. There is also a close correspondence between the estimated profit rates on American gold imports and British gold exports on each of the eight dates.

In table 2, the estimated profit rates computed with the weekly data for the period 1899–1908 appear. Out of the 518 weeks for which

---

4 8405 for October 1891. For the week September 18-October 3, 1891, Andrew reports an exchange rate of 4 84 each day. Instead of indicating a profit on American gold exports, Andrew's quotes indicate that it was profitable to import gold into the United States on October 3, 1891 Because of the disparity between exchange rate quotes on these dates, estimated profit rates for February 1, 1890, and October 3, 1891, are not reported in table 1.

[10] On each shipping date between 1890 and 1908, rates of return were computed for shipping times of 7–14 days. As shown in table 6, which will be discussed below, the estimated profits were sensitive to changes in the assumed shipping times. For the sake of brevity, only the estimates for 10 days shipping time are reported. These estimates were selected because the average transatlantic crossing reportedly took approximately 10 days during the time period being examined.

[11] The profit rates on British gold imports and American gold exports should, of course, be equal on each date. The slight differences observed are presumably attributable to rounding errors and a lack of precision in the data. To establish bounds on the estimates, both estimated profit rates are reported on each date in tables 1 and 2.

Table 1 —ESTIMATED PROFIT RATES ON GOLD SHIPMENTS ASSUMING 10 DAYS SHIPPING TIME AND USING MONTHLY DATA, 1890–98

Date	Import Profit (%)	Export Profit (%)
	Estimated Profits on British Gold Imports and U.S. Gold Exports	
August 1890*	2065	2068
March 1891	1151	1152
April 1891	.0730	0732
May 1891	1341	1342
June 1891	.0891	0893
April 1892	0377	0378
May 1892	0036	0037
June 1892	0540	.0541
July 1892	0540	0542
August 1892	0540	0541
September 1892	0026	0029
May 1893	0874	0874
June 1893*	3469	3477
March 1894	0748	0748
April 1894	0857	0857
May 1894*	1955	1955
June 1894	1567	1568
July 1894	0060	0061
August 1894	1636	1636
November 1894	0060	0061
January 1895	1636	1637
February 1895*	2619	2620
March 1895*	2534	2536
April 1895*	3574	3578
May 1895*	2629	2630
June 1895	0865	[illegible]
July 1895*	.3690	
August 1895*	5187	
September 1895*	4684	
November 1895	1537	
December 1895*	2023	
January 1896*	.3108	
February 1896	1054	
March 1896	0074	
April 1896	1601	
May 1896*	2003	
June 1896	1088	
July 1896	0608	[illegible]
August 1896*	3108	3114
May 1897	0016	0017
	Estimated Profits on U S Gold Imports and British Gold Exports	
January 1891	0145	0143
December 1891	0522	0521
November 1893	.0810	0809
October 1897	0004	0003
April 1898	1004	1004
May 1898	.0743	0744
October 1898	0589	0588

* Denotes dates when estimated gold arbitrage profits existed assuming a larger direct shipping cost of 0 5 percent during 1890–1904

TABLE 2

Estimated Profit Rates on Gold Shipments Assuming 10 Days Shipping Time and Using Weekly Data, 1899–1908

Date	Import Profit (%)	Export Profit (%)
	Estimated Profits on British Gold Imports and U.S. Gold Exports	
June 3, 1899	.0918	.0919
June 10, 1899	.0323	.0324
June 24, 1899	.0037	.0038
October 21, 1899	.0192	.0196
December 30, 1899	.0070	.0072
January 6, 1900	.0337	.0339
January 13, 1900	.1019	.1021
April 21, 1900	.0155	.0153
April 28, 1900	.0267	.0266
May 5, 1900	.0651	.0650
May 12, 1900	.1145	.1144
May 19, 1900	.0738	.0739
May 26, 1900	.0735	.0735
March 2, 1901	.0201	.0200
March 23, 1901	.0224	.0223
March 30, 1901	.0721	.0721
April 6, 1901	.0768	.0770
April 13, 1901	.0204	.0206
April 27, 1901	.0877	.0882
May 4, 1901	.0170	.0179
May 11, 1901	.0437	.0439
May 18, 1901	.0857	.0858
May 25, 1901	.1389	.1389
June 1, 1901	.1830	.1832
June 8, 1901	.1313	.1315
June 15, 1901	.1105	.1108
June 22, 1901	.0264	.0268
August 10, 1901	.0789	.0790
May 9, 1903	.0253	.0253
May 23, 1903	.0047	.0046
May 30, 1903	.0524	.0524
June 6, 1903	.0023	.0023
August 6, 1904	.0621	.0620
August 13, 1904	.0941	.0940
August 20, 1904	.0130	.0130
	Estimated Profits on U.S. Gold Imports and British Gold Exports	
October 13, 1900	.1115	.1113
October 27, 1900	.0018	.0016
November 14, 1903	.1415	.1413
November 28, 1903	.0611	.0607
December 5, 1903	.1421	.1419
December 12, 1903	.1585	.1584

TABLE 2 (*Continued*)

Date	U.K Import Profit (%)	U.S. Export Profit (%)
January 2, 1904	.0210	0210
May 5, 1906*	0556	0555
July 21, 1906*	0208	0208
September 8, 1906*	0662	0659
September 15, 1906*	0052	0051
September 22, 1906*	0973	0972
September 29, 1906*	0501	0499
December 22, 1906*	0081	0078
March 23, 1907*	0989	.0988
March 30, 1907*	1101	1101

* Denotes dates when estimated gold arbitrage profits existed assuming direct shipping costs of 0.5 percent during 1890–1904 and 0.3125 percent during 1905–8

returns could be estimated,[12] there were 35 weeks when it was profitable to ship gold from the United States to Britain and 16 weeks when it was profitable to ship gold in the opposite direction. The maximum estimated profit rate on British gold imports was 0.18 percent on June 1, 1901, and the minimum was 0.002 percent on June 6, 1903. The range of the estimated profit rates on American gold imports was between 0.16 percent on December 12, 1903, and 0.002 percent on October 27, 1900. On each date in table 2, the estimated profit rates on imports and exports for gold flows in the same direction are very similar.

It is apparent from tables 1 and 2 that the
estimated profit opportunities for shippi
States to Britain than there were for ship
direction. However, there were no British
tunities after August 1904. This indicates
ally "weak" (i.e., $S_t > 4.8665$) relative to
the period 1890–1904. The weakness of the
March 1894–August 1896 is particularly eviden

---

[12] An observation of the spot exchange rate for the week ending on December 31, 1904, was not reported preventing the computation of returns for that date. The last date on which rates of return were estimated was December 12, 1908. This was because the last observation of the British weekly average call money rate was for December 25, 1908, making it impossible to approximate the British interest cost with eq. (7) after December 12, 1908.

[13] The "flight from the dollar" during this period usually is attributed to uncertainty about the maintenance of the gold standard. The proponents of "free silver" attained the peak of their political influence in July 1896 when William Jennings Bryan was nominated for president by the Democratic party (see Friedman and Schwartz 1963, pp. 104–19)

III. The Sensitivity of the Estimated Gold Shipping Profits

In Section II, I estimated rates of return on gold shipments on 626 dates between 1890 and 1908. Profits were indicated on 98 dates. Of these, 89 occurred between 1890 and 1904. The frequency and concentration of profit opportunities raise some questions about the reliability of the estimates.

It is possible that there are errors in the spot exchange rate and interest rate data sets. Errors in the data could produce spurious estimates of profit opportunities. Of course, errors in the data could also prevent the identification of true profit opportunities. Without more reliable data sets, it is difficult to determine the extent to which the estimates of profits on gold arbitrage operations were affected by data errors.

The fairly large number of estimated profit opportunities could be due to systematic underestimates of the costs of shipping gold. As mentioned in Section I, commission expenses on currency conversions and securities transactions were ignored. The omission of these costs biases estimated profit rates upward. Another possible source of systematic error could be the use of weekly average call money rates to approximate the interest costs of gold shipments. If the weekly average call money rates were systematically less than true interest costs, this would also bias the estimated profit rates upward.

However, systematic underestimates of the cost of shipping gold would tend to produce estimated profit opportunities uniformly throughout the period 1890–1908. Instead, the estimated profit opportunities reported in tables 1 and 2 are highly concentrated in the years 1890–1904. This concentration suggests a systematic underestimate of costs that was declining over the sample period. One possibility is that, while shipping speeds were gradually increasing during the sample period, the average time for a transatlantic crossing exceeded the assumed 10 days prior to 1905. As can be seen in table 3 (which reports the frequencies of profit opportunities as the assumed shipping time varies between 7 and 14 days), the frequencies of estimated profits are somewhat sensitive to changes in the number of days assumed to complete a gold arbitrage operation. If it is assumed that the average transatlantic crossing time was 11 days prior to 1905, then the number of estimated profit opportunities that could have occurred during 1890–1904 drops from 89 to 83.

It could also be the case that the estimate of the direct shipping expenses used was too low during the early part of the sample period. The estimate of 0.3125 percent was made in 1906. As noted by the source of this estimate (Myers 1931, p. 342), direct shipping costs were declining over time.

TABLE 3

FREQUENCIES OF VIOLATIONS OF THE ESTIMATED GOLD POINTS WHEN THE ASSUMED SHIPPING TIME VARIES BETWEEN 7 AND 14 DAYS, 1890–1908

	Number of Profit Opportunities for Various Assumed Shipping Times							
	7 Days	8 Days	9 Days	10 Days	11 Days	12 Days	13 Days	14 Days
A. Monthly data, 1890–98:								
U.K. imports/U.S. exports	40	40	40	40	39	38	38	37
U.S. imports/U.K. exports	12	10	8	7	6	5	5	5
Totals	52	50	48	47	45	43	43	42
B. Weekly data, 1899–1908:								
U.K. imports/U.S. exports	48	45	43	35	32	30	25	23
U.S. imports/U.K. exports	20	19	17	16	13	13	13	11
Totals	68	64	60	51	45	43	38	34
C. Combined weekly and monthly data, 1890–1908:								
U.K. imports/U.S. exports	88	85	83	75	71	68	63	60
U.S. imports/U.K. exports	32	29	25	23	19	18	18	16
Totals	120	114	108	98	90	86	81	76

If direct shipping costs are assumed to be 0.005 during 1890–1904 and 0.003125 during 1905–8, then the total number of dates when estimated profits existed drops from 98 to 23. In addition, the remaining profit opportunities are more uniform [illegible] the 1890–1904 and 1905–8 subperiods. This [illegible] ining tables 1 and 2. The 23 dates on which es[illegible] with the revised direct shipping cost assum[illegible] asterisks.[11] There are 14 remaining estima[illegible] between 1890 and 1904 and nine between 1[illegible] gests that the frequency of gold point violation[illegible] and their concentration in the early years of the sam[illegible] be attributable to a systematic underestimate of the cost of shipping gold during the 1890–1904 subperiod.

The use of possibly biased estimates of gold arbitrage profits could lead to incorrect inferences regarding the efficiency of the gold standard system. To reduce this danger, I will use two sets of estimated gold point violations in the appraisal of market efficiency in Section IV. The first set is those obtained with the original direct shipping cost assumption (i.e., a constant 0.3125 percent throughout the 1890–1908 period). This set will be referred to as the "original" estimates. The second set is those obtained with the revised direct shipping cost assumption (i.e., 0.5 percent during 1890–1904 and 0.3125 percent thereafter). The second set will be referred to as the "revised" estimates. The examination of the revised estimates should yield more reliable inferences regarding the efficiency of the gold standard system.

---

[11] Revised profit estimates were obtained by subtracting 0.1875 percent (the difference between 0.5 percent and 0.3125 percent) from all of the original estimated profits before 1905 listed in tables 1 and 2.

IV. Did the Gold Standard System Function Efficiently?

Many criteria can be used to judge the efficiency of the gold standard. I consider only two in this study. The first is how rapidly gold point violations disappeared. The second is whether gold always moved in the profitable direction whenever a gold point violation occurred.

Aside from the time required to transport gold between countries, one would expect international gold flows to eliminate profit opportunities fairly quickly. If this was the case, gold point violations should have occurred intermittently rather than in sequences on consecutive dates.

In the original estimated profit opportunities in tables 1 and 2, there were many prolonged sequences of gold point violations. In table 1, it apparently was profitable to ship gold from the United States to Great Britain during the following consecutive months: March–May 1891, April–September 1892, May–June 1893, March–August 1894, January–September 1895, and November 1895–July 1896. It also apparently was profitable to ship gold from Britain to the United States in April and May 1898. In table 2, there are several instances of multiple gold point violations that occurred within a month or a group of consecutive months. For example, the British gold import point was violated in 14 out of 17 weeks between March and June 1901. The persistence of gold point violations observed in tables 1 and 2 suggests that the gold standard system did not always adjust rapidly.

Looking at the dates of the revised profit opportunities, which are marked with asterisks in tables 1 and 2, there still remain instances where arbitrage opportunities are clustered together on consecutive dates. For example, in table 1, it apparently was profitable for Americans to export gold in 9 out of the 12 months between February 1895 and January 1896. Similarly, in table 2, it apparently was profitable for the British to export gold during the four consecutive weeks between September 8 and September 29, 1906. Thus, even when only the revised profit opportunities are examined, it does not appear that the gold market was always able to eliminate profit opportunities quickly.

Did gold always flow in the profitable directions? Having identified the dates on which estimated gold shipment profit opportunities occurred, one can see whether gold flowed in the predicted directions by examining the monthly net gold export figures of the United States and Great Britain.[15]

A minor complication arises in the identification of the months affected by some of the estimated profit opportunities because of the length of time required to ship gold between the two countries. Since the average time for a transatlantic crossing was 10 days, gold shipped from one country near the end of one month could arrive in the second country in the early part of the next. Thus, while a profit opportunity should affect the net gold export figures of the exporting country in the same month the profit opportunity occurred, the net gold export figures of the importing country possibly would not be affected until the month following the profit opportunity.[16]

---

[15] The monthly net gold export figures for each country were obtained from the N.B.E.R. The British figures are from Business Cycle Series 14.108 "Great Britain Excess of Gold Exports over Gold Imports, Total United Kingdom." The American figures are from Business Cycle Series 14,112, "United States, Excess of Gold Exports over Gold Imports." The figures examined here are the net gold exports from one country to the "rest of the world," not just restricted measures of gold flows between the United States and Great Britain.

[16] In this study, the possibility of a gold shipment affecting one country's gold exports in one month and the other country's gold imports in the next month arose only during 1899–1908 when weekly data were used. During 1890–98
used to estimate profit opportunities. Since each monthly
later than the seventh day of each month, gold shipped
observation should have reached the other side of the
the month. While avoiding one source of ambiguity, the
another potential problem in interpreting gold flows
opportunities. Since profit opportunities were estimated
each month during 1890–98, gold could appear to flow
some months because of rapid and drastic changes in
remaining weeks of those months. For example, if
sufficiently volatile, the American gold export point could
week of a given month and then the American gold import point
one or more of the remaining weeks of that month. The monthly data used in this study would identify only the violation of the American gold export point leading to the prediction of net gold exports by the United States in that month. If subsequent gold imports in response to the unobserved violations of the American gold import point swamped the initial gold exports, the United States would be a net gold importer in a month when gold exports were predicted. Given the hypothetical changes in market conditions during the month, it would be incorrect to conclude that gold flowed in the unprofitable direction and that the market was inefficient. Without weekly data, it is impossible to be sure that any apparently unprofitable gold flow observed during 1890–98 did not occur simply because of such a violent within month exchange rate reversal. However, examining the estimated gold point violations in table 2, one can see that no such reversal occurred within any month during 1899–1908 when weekly data were used. Unless foreign exchange markets were much less volatile during 1899–1908, it seems unlikely that gold point violations in opposite directions occurred in any month during 1890–98.

To deal with this problem, I assumed that all gold shipment operations required 10 days to complete. If a profit opportunity occurred prior to the last 10 days of a month, I assumed that it affected the net gold export figures of both the exporter and the importer in the same month. (For example, as seen in the first col. of table 2, the estimated profit opportunity that occurred on May 19, 1900, should have triggered American gold exports on that date leading to British gold imports on May 29, 1900.) If a profit opportunity arose during the last 10 days of a month, I assumed that it affected the net gold export figure of the exporting nation in that month and the net gold export figure of the importing nation in the following month. (For example, again as seen in the first col. of table 2, the May 26, 1900, estimated profit opportunity should have resulted in American gold exports on that date and British gold imports on June 5, 1900.) In a month in which more than one estimated profit opportunity occurred, it was possible that the net gold import figures of the importing nation were affected in 2 months. (For example, the profit opportunities occurring on May 19 and 26, 1900, were assumed to produce British gold imports in both May and June 1900 while producing American gold exports only in May 1900.)

If the gold standard system was functioning efficiently, the United States should have exported gold and/or Great Britain should have imported gold whenever it was profitable to ship gold from the United States to Britain. Similarly, if it was profitable to ship gold from Britain to the United States, Britain should have exported gold and/or the United States should have imported gold. Either the exporting of gold when gold imports were profitable or the importing of gold when gold exports were profitable was inconsistent with the efficient functioning of the system. The validity of this second criterion for market efficiency is demonstrated in Appendix B below.

In table 4, the net gold exports of the United States and Great Britain in the months when gold shipments from the United States to Britain were estimated to be profitable are presented. If gold was flowing in the profitable direction, the United States should have been a net gold exporter (indicated by positive net gold exports) and/or Britain should have been a net gold importer (indicated by negative net gold exports) in these months. While the British and American net gold export figures are usually consistent with these predictions, there are several months in which gold was apparently moving in an unprofitable direction.

Table 5 presents British and American net gold export figures in the months when profits appeared to exist on gold shipments from Britain to the United States. In these months, Britain should have been a net gold exporter and/or the United States should have been a

TABLE 4

American and British Net Gold Exports in Months When Gold Shipments from the United States to Great Britain Were Estimated to Be Profitable, 1890–1908

Month Gold Shipped from U.S.	U.S. Net Gold Exports ($ Thousands)	Month Gold Arrived in U.K. (10 Days Shipping Time)	U.K. Net Gold Exports (£ Thousands)
August 1890	411*	August 1890	–2,412*
March 1891	4,542	March 1891	–140
April 1891	13,930	April 1891	95
May 1891	30,368	May 1891	–4,740
June 1891	15,539	June 1891	–5,239
April 1892	7,035	April 1892	–989
May 1892	3,263	May 1892	–1,178
June 1892	16,635	June 1892	–1,307
July 1892	10,240	July 1892	–1,040
August 1892	5,717	August 1892	–1,758
September 1892	2,324	September 1892	–486
May 1893	15,206	May 1893	–953
June 1893	1,702*	June 1893	–3,852*
March 1894	2,883	March 1894	–1,367
April 1894	9,400	April 1894	–2,245
May 1894	23,097*	May 1894	–4,588*
June 1894	22,298	June 1894	2,952
July 1894	12,752	July 1894	–1,196
August 1894	1,841	August 1894	–1,918
November 1894	–1,551	November 1894	1,253
January 1895	24,906	January 1895	–742
February 1895	–4,142*	February 1895	789*
March 1895	–4,227*	March 1895	
April 1895	–2,177*	April 1895	
May 1895	–3,463*	May 1895	
June 1895	–2,164	June 1895	
July 1895	3,124*	July 1895	
August 1895	15,007*	August 1895	
September 1895	16,507*	September 1895	
November 1895	13,256	November 1895	
December 1895	14,068*	December 1895	
January 1896	25*	January 1896	
February 1896	–9,468	February 1896	
March 1896	–355	March 1896	[illegible]
April 1896	2,513	April 1896	404
May 1896	18,399*	May 1896	–994*
June 1896	5,934	June 1896	–1,390
July 1896	10,263	July 1896	1,456
August 1896	–2,317*	August 1896	914*
May 1897	8,517	May 1897	–1,237
June 1899†	17,803	June 1899	–2,188
		July 1899	–3,604
October 1899	–8,163	October 1899	139
December 1899‡	6,237		
January 1900	3,699	January 1900	–3,878
April 1900‡	–1,427		

TABLE 4 (*Continued*)

Month Gold Shipped from U.S.	U.S. Net Gold Exports ($ Thousands)	Month Gold Arrived in U.K. (10 Days Shipping Time)	U.K. Net Gold Exports (£ Thousands)
May 1900†	8,526	May 1900	−1,378
		June 1900	428
March 1901†	−2,030	March 1901	−958
April 1901†	2,668	April 1901	−2,186
May 1901†	8,328	May 1901	−1,528
June 1901†	2,084	June 1901	−955
		July 1901	−911
August 1901	−3,340	August 1901	1,825
May 1903†	13,025	May 1903	−1,045
June 1903	9,740	June 1903	−1,262
August 1904	2,998	August 1904	−1,242

SOURCES.—Net gold export figures were obtained from N.B.E.R. The British figures are from Business Cycle Series 14,108. The American figures are from Business Cycle Series 14,112.

* Denotes net gold exports in months when estimated gold arbitrage profits existed assuming direct shipping costs of 0.5 percent during 1890–1904 and 0.3125 percent during 1905–8.

† Because of multiple dates of profit opportunities during this month, gold shipments from the United States are assumed to have arrived in Britain during both this month and the following month.

‡ Gold shipments from the United States during this month are assumed to have arrived in Britain only during the following month.

net gold importer. During each month listed in table 5, the United States was a net gold importer, but Britain was not always a net gold exporter.

Table 6 provides a summary of the correspondence between the directions of actual and predicted gold flows from tables 4 and 5. It is apparent that gold imports usually occurred when estimated gold import profits existed. The United States was a net gold importer in all 18 months in which gold was predicted to flow from Britain to the United States. Britain was a net gold importer in 51 out of the 55 months gold was predicted to flow from the United States to Britain. However, the four instances when Britain exported gold when imports were predicted are not consistent with the hypothesis of gold market efficiency.

The striking feature of table 6 is that most of the discrepancies between actual and predicted gold flows arise when gold export profits apparently existed. Britain was a net gold exporter in only 8 out of the 16 months when it appeared to be profitable to ship gold from Britain to the United States. The United States was a net gold exporter in 41 out of the 54 months when gold flows from the United States to Britain were predicted. Gold imports during periods when gold exports were profitable are inconsistent with the hypothesis of gold market efficiency.

Actual gold flows in the months affected by the 23 revised es-

TABLE 5

British and American Net Gold Exports in Months When Gold Shipments from Great Britain to the United States Were Estimated to Be Profitable, 1890–1908

Month Gold Shipped from U.K.	U.K. Net Gold Exports (£ Thousands)	Month Gold Arrived in U.S. (10 Days Shipping Time)	U.S. Net Gold Exports ($ Thousands)
January 1891	264	January 1891	−670
December 1891	300	December 1891	−5,764
November 1893	871	November 1893	−4,140
October 1897	2,013	October 1897	−11,462
April 1898	−2,447	April 1898	−31,256
May 1898	−6,642	May 1898	−13,213
October 1898	−743	October 1898	−15,458
October 1900†	1,324	October 1900	−10,289
		November 1900	−11,965
November 1903†	1,996	November 1903	−10,378
December 1903	733	December 1903	−15,766
January 1904	168	January 1904	7,634
May 1906	−360*	May 1906	29,189*
July 1906	−1,207*	July 1906	8,532*
September 1906†	5,403*	September 1906	−29,152*
		October 1906	−20,176*
December 1906‡	1,008*		
		January 1907	820*
March 1907‡	−2,069*		
		April 1907	−2,755*

Sources.—Net gold export figures were obtained from N.B.E.R. The British figures are from Business Cycle Series 14.108. The American figures are from Business Cycle Series 14.112.

* Denotes net gold exports in months when estimated gold arbitrage profits ex
of 0.5 percent during 1890–1904 and 0.3125 percent during 1905–8.

† Because of multiple dates of profit opportunities during this month
to have arrived in the United States during both this month and the fo

‡ Gold shipments from Britain during this month are assumed to h
the following month.

TABLE 6

Correspondence between the Directions of Actua in Months Affected by Estimated Gold Shipment 1890–1908

	Great Britain	United States
Gold imports:		
Number of months gold imports actually occurred when gold imports were predicted	51	18
Total number of months gold imports were predicted	55	18
Gold exports:		
Number of months gold exports actually occurred when gold exports were predicted	8	41
Total number of months gold exports were predicted	16	54

TABLE 7

THE CORRESPONDENCE BETWEEN THE DIRECTIONS OF ACTUAL AND PREDICTED GOLD FLOWS IN MONTHS AFFECTED BY ESTIMATED GOLD SHIPMENT PROFIT OPPORTUNITIES ASSUMING DIRECT SHIPPING COSTS OF 0.5 PERCENT DURING 1890–1904 AND 0.3125 PERCENT DURING 1905–8

	Great Britain	United States
Gold imports:		
Number of months gold imports actually occurred when gold imports were predicted	13	6
Total number of months gold imports were predicted	14	6
Gold exports:		
Number of months gold exports actually occurred when gold exports were predicted	1	9
Total number of months gold exports were predicted	5	14

timated gold point violations are indicated by asterisks in tables 4 and 5. It can be seen that there are still instances when gold apparently flowed in unprofitable directions.

Table 7 provides a summary of the correspondence between actual and predicted gold flows using the revised direct shipping cost assumptions. Table 7 shows that most of the instances when gold apparently flowed in an unprofitable direction occurred when exports were predicted. When import profits were predicted, Britain was a net gold importer in 13 out of 14 months and the United States was a net gold importer in all 6 months. By contrast, when exports were predicted, Britain was a net gold exporter in only 1 out of 5 months, and the United States was a net gold exporter in 9 out of 14 months.

During the years 1899–1908, weekly data on exports and imports of gold at New York City are available.[17] These weekly data permit a somewhat more precise check on the degree of correspondence between actual and predicted gold flows than was possible with the monthly gold flow data.

In table 8, the exports and imports of gold at New York City in weeks when gold shipments from the United States to Great Britain were estimated to be profitable are presented. In most of these weeks, New York City experienced large net gold exports. However, there are several weeks when net gold exports were negative.

---

[17] The weekly gold export and import figures for the port of New York were obtained from Andrew (1910, pp. 172–76, table 11). The original source of these data was the *Commercial and Financial Chronicle*.

TABLE 8

EXPORTS AND IMPORTS OF GOLD AT NEW YORK CITY IN WEEKS WHEN GOLD SHIPMENTS FROM THE UNITED STATES TO GREAT BRITAIN WERE ESTIMATED TO BE PROFITABLE, 1899–1908
($ Thousands)

Date	Exports	Imports	Net Exports	Date	Exports	Imports	Net Exports
June 3, 1899	1,264	42	1,222	April 27, 1901	560	18	542
June 10, 1899	4,545	768	3,777	May 4, 1901	4,018	68	3,950
June 24, 1899	6,324	51	6,273	May 11, 1901	760	24	736
October 21, 1899	27	110	−83	May 18, 1901	2,199	33	2,166
December 30, 1899	4,915	54	4,861	May 25, 1901	766	31	735
January 6, 1900	2,938	10	2,928	June 1, 1901	4,071	30	4,041
January 13, 1900	1,046	102	944	June 8, 1901	7	37	−30
April 21, 1900	15	13	2	June 15, 1901	426	47	379
April 28, 1900	1,110	69	1,041	June 22, 1901	4,523	42	4,481
May 5, 1900	2,692	25	2,667	August 10, 1901	12	67	−56
May 12, 1900	2,909	7	2,902	May 9, 1903	1,430	207	1,223
May 19, 1900	3,231		3,207	May 23, 1903	3,607	48	3,560
May 26, 1900	1,479		1,156	May 30, 1903	3,491	84	3,408
March 2, 1901	3		66	June 6, 1903	5,038	5	5,033
March 23, 1901	7		39	August 6, 1904	7	30	−23
March 30, 1901	383		263	August 13, 1904	3,008	9	3,000
April 6, 1901	1,543		[illegible]19	August 20, 1904	3,054	74	2,980
April 13, 1901	28		0*				

SOURCE.—Andrew (1910) pp. 172–76, table 1

* Net exports of gold from New York during [illegible] $200

Table 9 presents gold export and import figures for New York City 2 weeks after an estimated profit opportunity on gold shipments from Britain to the United States arose. (The 2-week lag was again employed to allow for shipping time.) In all 16 cases, New York City experienced large net gold inflows following estimated violations of the American gold import point. As indicated by asterisks, these cases include nine instances when violations of the American gold import point occurred using the revised direct shipping cost assumptions.

Table 10 provides a summary of the information reported in tables 8 and 9. The only discrepancies between the directions of actual and predicted gold flows occurred when gold exports were predicted. Gold appears to have flowed in an unprofitable direction in six out of the 35 cases where estimated violations of the American gold export point occurred using the original direct shipping cost assumptions. During the period 1899–1908, there was no instance when an estimated violation of the American gold export point occurred using the revised direct shipping cost assumptions.

The examination of the weekly gold flow data for New York City corroborates the findings based on the monthly net gold export figures for Britain and the United States. On occasion, gold apparently flowed in unprofitable directions, and such unprofitable gold flows usually occurred when gold exports were predicted.

Based on the evidence presented in tables 4–10, it appears that the gold standard system was inefficient. Gold point violations seemed to persist for extended periods, and gold sometimes flowed in unprofitable directions. A partial explanation for these indications of inefficiency is presented in the next section.

V. Government Intervention in the Gold Market

One explanation for the apparent inefficiency of the gold standard system is interference in the workings of the system by governmental bodies or their agents. The principal goal of monetary policy in each nation under the gold standard was the maintenance of convertibility between its currency and gold at the legal rate of exchange. Gold exports drained reserves from a nation's monetary authority. If the monetary authority had a legally required or self-imposed minimum ratio of reserves to liabilities, protracted gold outflows that lowered the reserve ratio could force the monetary authority to take actions to stem gold exports. The monetary authority was not subject to a corresponding pressure to curb continued gold imports as long as it was willing to add to its reserve holdings in the form of gold. Because of this asymmetry, governmental bodies were more likely to attempt to prevent or limit gold exports than gold imports. The preponderance

TABLE 9

Exports and Imports of Gold at New York City Two Weeks after the Dates When Gold Shipments from Great Britain to the United States Were Estimated to Be Profitable, 1899–1908

($ Thousands)

Date of Gold Point Violation	Week Gold Arrived in U.S. (10 Days Shipping Time)	Exports	Imports	Net Exports
October 13, 1900	October 27, 1900	13	3,395	−3,382
October 27, 1900	November 10, 1900	10	2,838	−2,828
November 14, 1903	November 28, 1903	55	4,051	−3,996
November 28, 1903	December 12, 1903	107	1,199	−1,092
December 5, 1903	December 19, 1903	111	2,504	−2,393
December 12, 1903	December 26, 1903	14	3,305	−3,291
January 2, 1904	January 16, 1904	95	3,073	−2,978
May 5, 1906	May 19, 1906	6	15,194	−15,188*
July 21, 1906	August 4, 1906	22	2,080	−2,058*
September 8, 1906	September 22, 1906	0	13,891	−13,891*
September 15, 1906	September 29, 1906	1	8,434	−8,433*
September 22, 1906	October 6, 1906	1	5,470	−5,469*
September 29, 1906	October 13, 1906	2	5,566	−5,564*
December 22, 1906	January 5, 1907	6	109	−103*
March 23, 1907	April 6, 1907	34	1,558	1,524*
March 30, 1907	April 13, 1907	0	432	−432*

Source.—Andrew (1910, pp. 172–76, table 11).

* Denotes net gold exports in weeks following gold point violations assuming direct shipping costs of 0.5 percent prior to 1905 and 0.3125 percent during 1905–8.

TABLE 10

The Correspondence between the Directions of
Flows at New York City in Weeks Affected by Bo
Estimates of Gold Shipment Profit Oppor

	Original Estim	
Gold imports		
Number of weeks gold imports actually occurred when gold imports were predicted	16	
Total number of weeks gold imports were predicted	16	9
Gold exports.		
Number of weeks gold exports actually occurred when gold exports were predicted	29	0
Total number of weeks gold exports were predicted	35	0

* The revised gold shipment profit estimates assumed direct shipping costs of 0.5 percent prior to 1905 and 0.3125 percent during 1905–8.

of instances in which gold was imported when gold exports were apparently profitable is consistent with the explanation that government interference was a source of the inefficiency of the gold market. However, it must be admitted that the assumed motivation for government interference does not lead to an explanation of why Britain exported gold on some occasions when it was apparently profitable to import gold from the United States.

How did governmental bodies attempt to control international gold flows?[18] Bloomfield (1959, p. 53) discussed several devices employed by the Bank of England to manipulate the gold points. To discourage British gold exports, the Bank occasionally imposed a premium on its selling price for gold bars and foreign gold coins. At other times, the Bank met demands for gold for export by payment in underweight coins. The effect of both of these measures was to lower the British gold export point (expressed in units of foreign currency per pound). On other occasions, the Bank of England tried to encourage British gold imports by lowering the gold import point. This was done by raising the Bank's purchase price for gold or by granting interest-free loans to gold importers during the time the gold was in transit.

The argument that government interference was a source of gold market inefficiency would be strengthened if it could be shown that some of the instances in which gold apparently moved in unprofitable directions occurred when the Bank of England was known to be tampering with the gold points. Unfortunately, Bloomfield did not specify the dates on which the Bank employed any of the techniques he described. However, some dates on which the U.S. Treasury or its agents intervened in the gold market are known, and there is a correspondence between these dates and apparently unprofitable gold movements.

Between April 14 and May 29, 1906, the U.S. Treasury took steps to eliminate the interest cost of gold imports into the United States. Initially, Secretary of the Treasury Leslie M. Shaw permitted one bank to count gold in transit to New York as part of its reserves. However, as this was blatantly illegal, Shaw next offered to place public deposits in non-interest-paying bank accounts equal in amount to any contracts the banks made for the importation of gold. These deposits had to be returned to the Treasury when the gold arrived, but, because the banks had the use of the funds deposited during the time the gold was in transit, the interest cost was eliminated. Secretary Shaw repeated his use of public deposits to stimulate gold imports between September 10 and October 23, 1906.[19]

---

[18] It is part of the gold standard lore that the Bank of England could control British gold flows by means of its discount policy. To halt British gold outflows, the Bank would raise the bank rate (i.e., its discount rate) while simultaneously draining reserves from the banking system. To stimulate British gold exports, the Bank would lower the bank rate while increasing the reserves available to the banking system. However, if the world's financial markets were integrated, it is doubtful that the Bank of England possessed such power. McCloskey and Zecher (1976) presented evidence suggesting that the traditional view of the gold standard era greatly exaggerated the ability of the Bank of England to direct international gold flows by varying the bank rate.

Secretary Shaw's measures perhaps had some effect. In 1906, the United States had net gold imports of approximately $12.5 million in April, $29.2 million in May, $29.2 million in September, and $20.2 million in October. However, it should be noted in table 2 that there were occasions in May and September 1906 when it was apparently profitable to import gold into the United States even when full interest costs had to be borne.

In addition to its own overt actions, the Treasury also condoned private attempts to affect gold flows. During 1895, a syndicate of bankers, headed by J. P. Morgan and August Belmont, engaged in activities designed to raise gold for the U.S. Treasury and to prevent gold exports from the United States. Between February and June 1895, the syndicate's actions were apparently successful. As seen in tables 1 and 4, American gold exports were apparently profitable in all 5 months, but the United States experienced net gold imports in each month. The bankers' syndicate collapsed in July 1895, and massive net exports of gold occurred through the end of the year. Also, following the nomination of William Jennings Bryan for president by the Democratic party, a syndicate of foreign exchange and gold shipping houses committed themselves to blocking gold outflows in the latter part of July 1896. This syndicate functioned
August. Its effectiveness is indicated by the fac
was a net gold importer during August 1896
in table 1 of a fairly substantial profit on go
month.[20]

Thus, in 6 of the months when the Unite
when gold exports were estimated to be profita

---

[19] Secretary Shaw's actions in 1906 are described in U.S. Secretary of the Treasury (1906, p. 39), Myers (1931, p. 343), Friedman and Schwartz (1963, pp. 155–56, n. 31) and Goodhart (1969, pp. 112–13).

[20] The Morgan-Belmont syndicate contracted to raise 3.5 million ounces of gold for the U.S. Treasury by the sale of Treasury bonds. The sale of some of those bonds in Europe accounted for part of the gold flows into the United States during February–June 1895. The syndicate also took steps to block gold exports by selling drafts on London at below market prices and not covering those sales by gold exports. To obtain the cooperation of other banks, the syndicate granted them shares in the underwriting in exchange for their pledges not to ship gold and to sell no foreign exchange at rates that would violate the gold export point. The underwriting profits were apparently adequate to more than cover any losses suffered by the syndicate in blocking profitable gold exports. Additional information about this syndicate and the one that operated in 1896 can be found in Friedman and Schwartz (1963, pp. 111–13, including n. 35).

condoned by the U.S. Treasury were known to be operating to promote American gold imports and block American gold exports. This observation provides support for the explanation that government interference was a source of the inefficiency of the gold standard system.

VI. Conclusions

This study has been concerned with violations of the gold points during the years 1890–1908. Dates on which arbitrage profits on gold shipments between the United States and Great Britain apparently existed were identified. The estimated gold point violations were used to assess two aspects of the efficiency of the gold standard: the speed with which the system adjusted to gold point violations and the direction of international gold flows in response to profit opportunities.

It was found that gold point violations often persisted for several successive months. This suggests that the gold standard system did not always eliminate profit opportunities quickly. It was also learned that gold sometimes flowed in apparently unprofitable directions in months when estimated gold point violations occurred. These findings are inconsistent with the view that the gold standard system functioned efficiently.

It was argued that a likely source of many of the observed inefficiencies was government interference with the workings of the system. Monetary authorities had incentives to restrict gold exports and to promote gold imports. They, unlike private market participants, were also willing to absorb any losses necessary to stimulate unprofitable gold flows. Evidence consistent with this argument was discovered: almost all instances where gold apparently flowed in unprofitable directions occurred when gold exports were predicted. It appears that the Bank of England and the U.S. Treasury occasionally tampered with the free workings of the gold market. In the short run at least, they apparently had some success in thwarting gold exports and stimulating gold imports.

The findings of this study have implications for the current debate concerning the desirability of restoring some form of gold standard. Advocates of a gold standard want to set limits on the discretionary power of central banks. The gold standard in effect in the late nineteenth and early twentieth centuries established such limits in the long run. However, as this study suggests, monetary authorities still possessed some degree of discretionary power in the short run. As long as central banks were willing to bear losses, they could delay the workings of the gold standard system. Presumably, central banks today would also be tempted to meddle with a similar system. If the goal of monetary reform is to eliminate all discretionary monetary power, something other than just the revival of the traditional gold standard is required.

Appendix A

The Effects of Currency Premiums in the United States on the Gold Points

The emergence of "currency premiums" in the United States disrupted the usual workings of the gold market on two occasions between 1890 and 1908 During the financial panics of 1893 and 1907, banks in the United States suspended the convertibility of deposits into currency in an attempt to retain their reserves As a result, currency went to a premium over certified checks. Between August 3 and September 2, 1893, and again between October 31 and December 28, 1907, currency was quoted at premiums of up to 4 percent.[21]

The effect of a currency premium was to alter the normal gold points Because gold arriving in New York could be converted into deposit balances for an additional gain, a premium on currency gave an added incentive to American gold imports. Let p_{t+k} denote the currency premium in New York on day $t + k$. With such a premium in effect, gold shipped from Britain on day t and arriving in New York k days later yielded a profit if the following inequality was satisfied.

$$\frac{(1 + p_{t+k})X}{S_t Y} > 1 + i_{US,t} + s \tag{A1}$$

Likewise, the profitability conditions for British gold exports to the United States shipped on day t and arriving on day $t + k$ became

$$\frac{(1 + p_{t+k})X}{F_t Y} > 1 + i_{UK,t} + s$$

The existence of currency premiums contribu gold into the United States during both panics In net gold imports were $10.6 million. The net impo States were $63.0 million in November 1907 and 1907. In terms of the estimated profits on gold sh mium of 1 percent[22] made it profitable to ship United States in August 1893. With 10 days shipping profit rate including the currency premium was 0.984 percent on American gold imports. During the panic of 1907, consideration of the currency premiums that could have been realized 10 days after each shipping date made it profitable to ship gold from Britain to the United States on October 26.

---

[21] In Sprague (1910), daily quotations of the currency premiums taken from New York newspapers are reported The quotes for the 1893 episode appear on p 187, and those for the panic of 1907 appear on pp 280–82.

[22] Gold shipped on August 5, 1893, and taking 10 days for delivery would have arrived in New York on August 15. On that date, Sprague (1910, p 187) reported a range of 1–1.5 percent in the bid quotes for currency Use of the minimum bid of 1 percent yields the most conservative estimates of gold shipment profits incorporating the currency premium

November 2, and November 9. Ten days after these three dates, the currency premiums were assumed to be 0.02, 0.035, and 0.015,[23] and the estimated profit rates on American gold imports were 0.826 percent, 2.81 percent, and 0.387 percent, respectively. In fact, the currency premiums during the panic of 1907 were sufficient to offset increases in the exchange rate up to $4.8875. As noted by Sprague (1910, p. 283), this was a level "well above the gold export point in normal times, and still gold imports continued on a massive scale."[24]

Appendix B

The Effects of Third Countries on Predicted Gold Flows between the United States and Great Britain

In Section IV above, I claimed that if it were profitable to ship gold from the United States to Great Britain, the United States should export gold and/or Britain should import gold. Likewise, if it were profitable to ship gold from Britain to the United States, Britain should export gold and/or the United States should import gold. However, the analysis of the workings of the gold standard conducted in Section I omitted consideration of other countries. What are the effects of this omission on predictions of the signs of British and American net gold exports given violations of the British-American gold points? For example, are there cases in which it is profitable to ship gold from the United States to Britain that, because of simultaneous violations of gold points with other countries, it is more profitable for the United States to import gold from another country and/or for Britain to export gold to another country? The purpose of this Appendix is to demonstrate that, under a given set of assumptions, such cases could not occur.

Assume that there are three countries: A, B, and C. To simplify the analysis, assume that there is only one currency and that the price of gold in each country and the costs of shipping gold between each pair of countries are expressed in terms of that currency. Let G_i represent the price of an ounce of gold in country i, and let c_{ij} denote the cost of shipping an ounce of gold between countries i and j.

The costs of international gold shipments are assumed to be increasing

---

[23] The currency premium used in each case was the lower bound of the daily range quoted by Sprague (1910, pp. 280–82).

[24] If the currency premiums that existed are ignored, it appeared to be profitable to export gold from the United States to Great Britain on November 9, 16, and 23, 1907. However, when the currency premiums are considered, shipments of gold from the United States on these dates would have resulted in significant losses. This is because the existence of currency premiums in the United States raised the American gold export and British gold import points. To ship gold from the United States to Britain on day t would require the initial conversion of deposit balances into currency at the premium in effect on day t. The currency premium p_t alters the American gold export and British gold import profitability conditions, respectively, in the following ways:

$$\frac{F_tY}{(1 + p_t)X} > 1 + i_{US,t} + s; \tag{A3}$$

$$\frac{S_tY}{(1 + p_t)X} > 1 + i_{UK,t} + s. \tag{A4}$$

functions of time. If it is further assumed that less time elapses when gold is shipped directly between any pair of countries (e.g., from A to B) than when it goes indirectly via a third country (e.g., from A to C to B), then the following restrictions must hold: $t_{AB} + t_{BC} > t_{AC}$; $t_{AB} + t_{AC} > t_{BC}$; $t_{AC} + t_{BC} > t_{AB}$.[25] These restrictions rule out the possibility that an equal or lower profit could be earned on a direct gold shipment between any two countries than could be earned on an indirect shipment between the two via a third country.

It is assumed that gold market participants in each country are profit maximizers and that at all times they know the price of gold in each country and the costs of shipping gold between countries. Suppose (as in the estimation of profit opportunities discussed in Secs. II and III above) that an "outsider" can observe only G_A, G_B, and t_{AB}. If $G_A < G_B - t_{AB}$, it is profitable to ship gold from A to B, and the outside observer would predict that A would export gold and B would import gold. However, since G_C, t_{AC}, and t_{BC} are unobserved, the outsider cannot be sure that a greater profit does not simultaneously exist that would lead to gold flows other than from A to B. Under the assumptions made here, there are six possible cases given that $G_A < G_B - t_{AB}$

CASE I: $G_A < G_B - t_{AB}$, and $G_B - t_{BC} \leqslant G_C \leqslant G_A + t_{AC}$.

In this case, the only gold point violation is $G_A < G_B - t_{AB}$. Market participants in A, B, and C will buy gold in A, ship it to B, and sell it there. The gold flow from A to B will continue until $G_A \geqslant G_B - t_{AB}$ and the profit opportunity is eliminated.

CASE II: $G_A < G_B - t_{AB}$, and $G_A + t_{AC} < G_C < G_B - t_{BC}$.

Under these conditions, it is profitable to ship gold from A to B, from A to C, and from C to B. However, the profit on shipping gold from A to B is greater than the profit on shipping it from A to C since $G_B - t_{AB} > G_B - t_{BC} - t_{AC} > G_C - t_{AC}$. Similarly, since $G_A + t_{AB} < G_A + t_{AC} + t_{BC} < G_C + t_{BC}$, the profit on shipping gold from A to B exceeds the profit on shipping it from C to B. Thus, since gold shipments from A to B produce the highest profit, market participants in A, B, and C will buy gold in A, ship it to B, and sell it there.[26]

CASE III: $G_A < G_B - t_{AB}$, and $G_C < G_A - t_{AC}$ [illegible]

Here, it is profitable to ship gold from A to B [illegible] A. However, the profit from C to B is greater [illegible] since $G_A - t_{AC} < G_B - t_{BC}$, and the profit from [illegible] profit from A to B since $G_C + t_{BC} < G_A + t_{AC}$ [illegible] gold will flow from C to B.

CASE IV. $G_A < G_B - t_{AB}$, and $G_C > G_B + t_{BC}$ [illegible]

---

[25] In addition to taking more time to complete, indirect gold shipments would also require more paperwork, inspections, and other handling. The additional charges for such services would also increase the cost of indirect gold shipments relative to that of direct gold shipments.

[26] It is useful to illustrate this case with a numerical example. Assume $G_A = \$230$, $G_B = \$300$, $G_C = \$275$, $t_{AB} = \$50$, $t_{AC} = \$40$, and $t_{BC} = \$20$. These values satisfy the conditions $G_A < G_B - t_{AB}$ and $G_A + t_{AC} < G_C < G_B - t_{BC}$ and the restrictions imposed on the shipping costs. The profit on shipping an ounce of gold from A to B is $20. The profit on shipping an ounce of gold from A to C or from C to B is only $5. Given that market participants in any country can buy an ounce of gold in A for $230, ship it to B for $50, and sell it in B for $300, shipments of gold from A to C or from C to B would result in opportunity losses of $15 per ounce shipped. Thus, as long as the conditions of this case hold, the only gold flow that will occur is that from A to B.

In this case, it is profitable to ship gold from A to B, from B to C, and from A to C. The profit from A to C is greater than the profit from B to C since $G_A + t_{AC} < G_B + t_{BC}$. The profit from A to C is greater than that from A to B since $G_C - t_{AC} > G_C - t_{BC} - t_{AB} > G_B - t_{AB}$. Thus, gold will flow from A to C.

Case V: $G_A < G_B - t_{AB}$, $G_A - t_{AC} \leqslant G_C$, and either $G_C < G_B - t_{BC} \leqslant G_A + t_{AC}$ or $G_C \leqslant G_A + t_{AC} < G_B - t_{BC}$.

Here, there are profit opportunities for gold shipments from A to B and from C to B. Gold will flow from A to B if $G_A + t_{AB} < G_C + t_{BC}$. Gold will flow from C to B if $G_A + t_{AB} > G_C + t_{BC}$. Gold will flow from both A and C to B if $G_A + t_{AB} = G_C + t_{BC}$.

Case VI: $G_A < G_B - t_{AB}$, $G_C \leqslant G_B + t_{BC}$, and either $G_A + t_{AC} < G_B - t_{BC} \leqslant G_C$ or $G_B - t_{BC} \leqslant G_A + t_{AC} < G_C$.

In this case, there are profit opportunities on gold shipments from A to B and from A to C. Gold will flow from A to B if $G_A - t_{AB} > G_C - t_{AC}$. Gold will flow from A to C if $G_B - t_{AB} < G_C - t_{AC}$. Gold will flow from A to both B and C if $G_B - t_{AB} = G_C - t_{AC}$.

There are three other cases that are conceivable. However, given the restrictions placed upon the shipping costs, these cases cannot occur.

Case VII. $G_A < G_B - t_{AB}$, and $G_B + t_{BC} < G_C \leqslant G_A + t_{AC}$

Profits are indicated on gold shipments from A to B and from B to C. These inequalities also imply that an indirect gold shipment from A to B to C is profitable while a direct shipment from A to C is not. These inequalities cannot be satisfied since they require $G_C \leqslant G_A + t_{AC} < G_A + t_{AB} + t_{BC} < G_B + t_{BC} < G_C$.

Case VIII: $G_A < G_B - t_{AB}$, and $G_B - t_{BC} \leqslant G_C < G_A - t_{AC}$.

Profits are indicated for gold shipments from A to B and from C to A. It is also possible to profit on an indirect gold shipment from C to A to B while a direct shipment from C to B is unprofitable. These inequalities cannot be satisfied because they require $G_C < G_A - t_{AC} < G_B - t_{AB} - t_{AC} < G_B - t_{BC} \leqslant G_C$

Case IX: $G_A < G_B - t_{AB}$, and $G_B + t_{BC} < G_C < G_A - t_{AC}$.

This might be termed the "merry-go-round case" since profits can be earned on gold shipments from A to B, from B to C, and from C to A. It cannot occur because it requires $G_C < G_A - t_{AC} < G_A + t_{AB} + t_{BC} < G_B + t_{BC} < G_C$.

In the first six cases that are possible, there are only three potential gold flows that maximize profits: from A to B, from A to C, and from C to B. Thus, when it is profitable to ship gold from A to B, A will export gold and/or B will import gold. Gold would be moving in unprofitable directions if either A were to import gold or B were to export gold when $G_A < G_B - t_{AB}$.

Based on these arguments, it does not appear that the lack of information about other countries undermines the second criterion for market efficiency employed in Section IV above. If it was profitable to ship gold from the United States to Great Britain, the United States should have exported gold and/or Britain should have imported gold. Likewise, if it was profitable to ship gold from Britain to the United States, Britain should have exported gold and/or the United States should have imported gold. If the gold standard system was functioning efficiently, gold should not have flowed into one of these countries when it was profitable for that country to export gold to the other, and gold should not have flowed out of one of these countries when it was profitable for that country to import gold from the other.

References

Andrew, A. Piatt. *Statistics for the United States, 1867–1909* Prepared for the National Monetary Commission. S. Doc. 570, 61st Cong., 2d sess Washington: Government Printing Office, 1910.

Bloomfield, Arthur I. *Monetary Policy under the International Gold Standard. 1880–1914.* New York: Federal Reserve Bank of New York, 1959

Borts, George H. Review of *International Financial Transactions and Business Cycles*, by Oskar Morgenstern. *J. American Statis. Assoc.* 55 (March 1960): 223–28.

Cohen, Benjamin J. *Balance-of-Payments Policy.* Harmondsworth: Penguin Books, 1969.

Friedman, Milton, and Schwartz, Anna Jacobson. *A Monetary History of the United States, 1867–1960* Princeton, N.J.: Princeton Univ. Press (for N B.E.R.), 1963.

Goodhart, Charles A. E. *The New York Money Market and the Finance of Trade, 1900–1913.* Cambridge, Mass.. Harvard Univ. Press, 1969.

McCloskey, Donald N., and Zecher, J Richard "How the Gold Standard Worked, 1880–1913." In *The Monetary Approach to the Balance of Payments*, edited by Jacob A. Frenkel and Harry G Johnson London Allen & Unwin, 1976.

Morgenstern, Oskar. *International Financial Transactions and Business Cycles* Princeton, N.J.: Princeton Univ. Press (for N B.E R.), 1959.

Myers, Margaret G. *The New York Money Market.* Vol. 1: *Origins and Development* New York: Columbia Univ. Press. 1931.

Sprague, Oliver M. W. *History of Crises under the National Banking System* Prepared for the National Monetary Commission S. Doc. 538, 61st Cong , 2d sess. Washington: Government Printing Office, 1910

U.S. Congress. Senate. National Monetary Commission *Statistics for Great Britain, Germany, and France, 1867–1909* S. Doc 578, 61st Cong , 2d sess Washington. Government Printing Office, 1910

U.S. Secretary of the Treasury *Annual Report of the Secretary of the Treasury on the State of the Finances for the Fiscal Year Ended June 30* [illegible]
59th Cong , 2d sess. Washington: Government [illegible]

An Equilibrium Model of Search Unemployment

James W. Albrecht
New York University

Bo Axell
Industrial Institute for Economic and Social Research

This paper develops a simple general equilibrium model with sequential search in which a nondegenerate wage offer distribution is endogenously determined. We use this model to analyze the comparative statics effects of increases in unemployment compensation on the unemployment rate and aggregate welfare taking into account the induced change in the wage offer distribution. Our results differ significantly from the predictions of the standard "partial-partial" model. For example, one can expect a selective increase in unemployment compensation, made available to those who impute a relatively low value to leisure, to decrease the equilibrium rate of unemployment.

I. Introduction

This paper develops a simple equilibrium model of search unemployment. By this term we mean the unemployment resulting from the rational rejection of available wage offers by unemployed job seekers in favor of further search for more lucrative offers. We focus on the effects of unemployment compensation on the equilibrium rate of search unemployment, both because unemployment compensation is an important policy issue and because the standard analysis of unem-

We thank Boyan Jovanovic for helpful comments. Albrecht's research was supported by the C. V. Starr Center for Applied Economics, New York University.

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0006$01.50

ployment compensation provides a convenient straw man against which to motivate our approach.

According to the standard search-theoretic model, an increase in unemployment compensation lowers the net cost of search to the unemployed, resulting in an increase in reservation wages and a consequent increase in the expected duration of search.[1] The key assumption underlying this model is that the wage offer distribution from which individuals search is exogenously given. This assumption is important for two reasons. First, we have no assurance that the idea of sequential search from a nondegenerate wage offer distribution makes any sense in equilibrium. There exists no simple general equilibrium model in which optimizing wage offers by firms combined with optimizing sequential search strategies by individuals result in a nondegenerate equilibrium wage offer distribution.[2] Second, even if one were to presume the existence of a nondegenerate equilibrium wage offer distribution, one would expect an increase in unemployment compensation to change that distribution. Since the standard comparative statics analysis of the effects of unemployment compensation is based on the notion of an exogenous and unchanging wage offer distribution, that analysis would seem to be of limited relevance (cf. Rosen 1977).

In this paper we develop a simple steady-state general equilibrium model with sequential search in which a nondegenerate wage offer distribution is endogenously determined. The search unemployment associated with the wage offer distribution is a Nash equilibrium outcome in the sense of being generated by the simultaneous optimizing behavior of firms and individuals in the eco
wage offer distribution and hence the eq
rate will vary with the amount of unemplo
able. We can therefore carry out a compa
effects of unemployment compensation t
dogeneity of the wage offer distribution

---

[1] The model underlying this result is presented, e.g., in
1976b). An example of the considerable empirical work that
predictions of the theory is Ehrenberg and Oaxaca (1976). The effects of imperfect "experience rating" in the system of financing unemployment compensation on the temporary layoff policies of firms have been stressed by Feldstein (1976) and are emphasized in the survey paper by Topel and Welch (1980). We do not deal with experience-rating issues in this paper.

[2] This is the well-known Rothschild (1973) criticism of the "partial-partial" nature of search theory. There are several models in the literature that feature nondegenerate endogenous wage (or price) offer distributions (e.g., Salop and Stiglitz 1977). However, almost all of these models are based on nonsequential, "noisy," or purely ad hoc search strategies; i.e., "dispersion equilibrium" is attained by sacrificing the tenet of sequential search. Three partial equilibrium models based on the sequential search strategy that generate dispersion equilibria are Axell (1977), Reinganum (1979), and Burdett and Mortensen (1980).

The basic ideas underlying our model are simple. We consider an economy in which a single good is produced with labor as the sole factor of production. Time is accounted in discrete periods, and in any period there are k individuals and n firms in the economy. Firms exist in perpetuity, but individuals suffer a constant "death risk" of τ ($0 < \tau < 1$) in the sense that an individual will exit the economy with probability τ at the end of any period. There is thus a flow of τk individuals into and out of the economy per period.

The product market is assumed to be an "auctioneer market" characterized by perfect information. Accordingly, all firms must offer the product for sale at a common price, which we normalize to unity. The labor market, on the other hand, is assumed to be a "search market" characterized by imperfect information in the standard search-theoretic sense of individuals knowing the form of the wage offer distribution but not knowing (prior to search) the identity of firms making particular offers.

The simplest way to generate "dispersion equilibria" in a model of this type is to allow for heterogeneity among individuals and/or firms. Our key assumption is that there are two types of individuals in the economy differing according to the value "imputed" to leisure. This assumption ensures that there can be at most two wages offered in equilibrium.

Assume for the moment the existence of the two-wage dispersion equilibrium. Let w_0 and w_1 denote the low and high wage, respectively, and let γ denote the fraction of firms offering the low wage. In equilibrium w_0 must be the reservation wage of those individuals who impute a low value to leisure, and w_1 must be the reservation wage of those who impute a high value to leisure. Individuals who place a low value on leisure will accept the first wage offer encountered, whereas those who place a high value on leisure will search until they encounter w_1. The amount of search in the economy—the unemployment rate—is therefore an increasing function of γ.

The requirement that w_0 and w_1 be reservation wages leads to two equilibrium conditions relating w_0, w_1, and γ. The third equilibrium condition is provided by the requirement that each firm make a profit-maximizing choice between w_0 and w_1. That some firms prefer to offer w_0 while others prefer to offer w_1 is ensured by allowing for heterogeneity among firms. Specifically, we assume a continuum of firms differing according to a "productivity index."[3]

---

[3] An alternative way to close the model is to assume homogeneous firms and impose an "equal profits" condition, i.e., to require that all firms be indifferent between offering w_0 and w_1. However, existence of the two-wage dispersion equilibrium becomes tenuous with this approach.

In the sections that follow we first specify the decision problems faced by utility-maximizing individuals and profit-maximizing firms, allowing us to derive the three equilibrium conditions relating w_0, w_1, and γ (Secs. II and III). In Section IV we establish a simple sufficient condition for the existence of the two-wage dispersion equilibrium, and in Section V we examine the comparative statics of increases in unemployment compensation. The effects of an increase in unemployment compensation will in general depend upon the distribution of the productivity index; however, for a broad class of distribution functions we show that a general increase in unemployment compensation increases the equilibrium unemployment rate, even when the endogeneity of the wage offer distribution is taken into account. However, a selective increase in unemployment compensation, given only to individuals who place a low value on leisure, will, for the same class of distributions, decrease the equilibrium unemployment rate. In Section VI we examine the efficiency implications of unemployment compensation. In our model, increases in unemployment compensation can enhance welfare by reallocating workers to more productive firms, even in the face of increasing unemployment, and we present a simple example to illustrate this effect. Finally, in Section VII we offer a concluding discussion.

II. Individuals

An individual entering the economy in any period is assumed to follow an optimal sequential search strategy with
imizing expected lifetime utility. His utility
to be of the form $u = x + vm$, that is,
consumption (x) and leisure (m). The para
sumption value" to leisure.

The variable m takes on the value of
whether the individual is working or search
variable x is given by the sum of consumptions ou
wage incomes. If the individual is working at a wage of w, then his wage income is w; otherwise his wage income is zero. Nonwage income consists of "dividends," θ, that is, the individual's per period share in economy-wide profits plus any unemployment compensation, b, received. Unemployment compensation is financed by lump-sum taxation out of dividends. Dividends are received whether the individual is working or searching,[1] whereas unemployment compen-

---

[1] Note that dividends therefore do not enter the individual's decision calculus. We take advantage of this to assume that unemployment compensation is financed out of dividends. We are therefore able to abstract from any effects brought about by the system of financing unemployment compensation.

sation is received only during periods of search. Individual utility in any period is thus given by

$$u = \begin{cases} w + \theta & \text{if working at a wage of } w \\ \theta + b + v & \text{if searching.} \end{cases} \tag{1}$$

The individual's search problem is as follows. When he enters the economy, he draws a wage of w at random from the wage offer distribution. If he accepts w, then he starts work immediately, supplying one unit of labor and consuming $w + \theta$ per period, and he continues to work at that wage for the duration of his lifetime. If, on the other hand, he rejects w, then he "consumes" an imputed leisure of v, receives a nonwage income of $\theta + b$, and with probability $1 - \tau$ survives to draw another wage at random from the wage offer distribution in the next period. Note that since an individual's lifetime is a "memoryless" random variable—that is, the "death risk" is constant—an individual who survives to draw another wage faces a decision problem identical to the one faced on entering the economy.

We assume two classes of individuals differing according to "leisure values." Let β denote the fraction of individuals with low values of leisure, v_0, and let $1 - \beta$ denote the fraction with high values of leisure, v_1. All individuals draw from the two-point wage offer distribution, drawing a wage of w_0 with probability γ and a wage of w_1 with probability $1 - \gamma$.

In order for (w_0, w_1, γ) to be an equilibrium distribution, w_0 must be the reservation wage of the v_0 individuals and w_1 must be the reservation wage of the v_1 individuals. If w_0 were less than the reservation wage of the v_0 individuals, then firms offering w_0 could attract no workers. If, on the other hand, w_0 were to exceed the reservation wage of the v_0 individuals, then any firm offering w_0 could reduce its wage offer without suffering any loss in labor supply. Likewise, to attract any v_1 individuals, w_1 must be no less than the reservation wage of that group. On the other hand, were w_1 to exceed the v_1 individuals' reservation wage, then w_1 could be reduced without any loss in labor supply.

These facts allow us to derive two equilibrium conditions relating w_0, w_1, and γ. Consider an individual with value of leisure v_0 who has drawn a wage of w_0. If he rejects w_0, then he enjoys a period of leisure valued at v_0, receives a nonwage income of $\theta + b$, and with probability $1 - \tau$ survives to draw another wage. The wage sampled on the subsequent draw equals w_1 with probability $1 - \gamma$; and if w_1 is in fact drawn, then that wage is accepted, leading to an expected future lifetime utility of $(w_1 + \theta)/\tau$. Otherwise the search process continues.

Thus, the value of further search, that is, the value of rejecting w_0, is

$$V^* = v_0 + b + \theta + (1 - \tau)\left[\frac{(1-\gamma)(w_1 + \theta)}{\tau} + \gamma V^*\right],$$

or

$$V^* = \frac{v_0 + b}{1 - \gamma(1 - \tau)} + \frac{(1-\gamma)(1-\tau)}{1 - \gamma(1-\tau)}\,\frac{w_1}{\tau} + \frac{\theta}{\tau}.$$

But if w_0 is the reservation wage for v_0 individuals, then $V^* = (w_0 + \theta)/\tau$, the value of accepting w_0. Thus we have

$$w_0 = \frac{(1-\gamma)(1-\tau)}{1 - \gamma(1-\tau)}\,w_1 + \frac{\tau(v_0 + b)}{1 - \gamma(1-\tau)} \tag{2}$$

The reservation wage property of w_1 is even simpler. The value of rejecting w_1 equals $(v_1 + b + \theta)/\tau$, whereas the value of accepting w_1 is $(w_1 + \theta)/\tau$. Thus,

$$w_1 - v_1 - b = 0;^5 \tag{3}$$

III. Firms

Firms are assumed to produce according to the linear production functions $y = \lambda l$, where the "productivity index" (output per worker) λ is distributed across firms according to the distr [illegible] As a normalization, we take the support o [illegible] denote the per period labor supply to a firm [illegible] w. Then the profits of a firm with productivi [illegible] its wage offer w are simply $\Pi(w; \lambda) = (\lambda -$ [illegible]

In a dispersion equilibrium, a fraction γ o [illegible] w_0 and a fraction $1 - \gamma$ offers w_1. (The conc [illegible] be defined below.) The requirement that firms [illegible] maximizing gives the final equilibrium condition relating w_0, w_1, and γ. The derivation of this equilibrium condition is illustrated in figure 1. First, firms with $\lambda \leq w_0$ do not operate; hence, only a fraction 1 $A(w_0)$ of all firms is "active." We assume that individuals search only

---

[5] In partial equilibrium search models a common necessary condition for the existence of a nondegenerate wage (or price) distribution is that the distribution of "search costs" not be bounded away from zero (see Axell 1977) In the general equilibrium context the forgone wage component of search cost is endogenously determined so that this necessary condition is automatically met

[6] We will assume that λ has a differentiable density function, $a(\lambda)$, whenever it is convenient to do so.

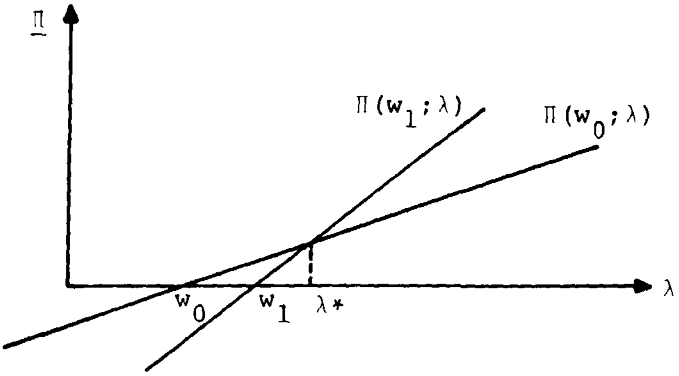


FIG. 1.—$\Pi(w_0; \lambda)$ and $\Pi(w_1, \lambda)$

from active firms. Next, let λ^* be defined by $\Pi(w_0; \lambda^*) = \Pi(w_1; \lambda^*)$; that is, λ^* is the productivity index such that a firm is indifferent between offering w_0 and w_1. Firms with $w_0 < \lambda \leq \lambda^*$ will offer w_0, while firms with $\lambda^* < \lambda \leq 1$ will offer w_1. Hence we have the equilibrium condition

$$\gamma = \frac{A(\lambda^*) - A(w_0)}{1 - A(w_0)}, \tag{4}$$

where

$$\lambda^* = \frac{w_1 l(w_1) - w_0 l(w_0)}{l(w_1) - l(w_0)}. \tag{5}$$

Finally, we need to derive $l(w_0)$ and $l(w_1)$. Consider a firm offering w_0. Only individuals with the low value of leisure, v_0, will accept this offer. There are $\tau k \beta$ v_0 individuals entering the economy per period, and if we let $\mu \equiv k/n[1 - A(w_0)]$, the ratio of individuals to active firms, then there are $\tau \mu \beta$ such individuals per active firm entering the economy each period. All of the $\tau \mu \beta$ v_0 individuals contacting a firm offering w_0 will accept that offer; hence, $l(w_0)$ can be computed as the sum of the $\tau \mu \beta$ individuals who accept w_0 in the current period, the $(1 - \tau)\tau\mu\beta$ surviving individuals who accepted w_0 in the previous period, and so on; that is,

$$l(w_0) = \tau\mu\beta[1 + (1 - \tau) + (1 - \tau)^2 + \ldots],$$

or

$$l(w_0) = \mu\beta. \tag{6}$$

Next, consider a firm offering w_1. All individuals contacting this firm accept w_1, and the number of contacts per firm per period is the sum of the

$\tau\mu\beta$	v_0 individuals entering the economy
$\tau\mu(1 - \beta)$	v_1 individuals entering the economy
$\tau\mu(1 - \beta)\gamma(1 - \tau)$	v_1 individuals who have searched once
$\tau\mu(1 - \beta)\gamma^2(1 - \tau)^2$	v_1 individuals who have searched twice

.

.

.

Thus, the number of acceptances per period is

$$\tau\mu\beta + \tau\mu(1 - \beta)[1 + \gamma(1 - \tau) + \gamma^2(1 - \tau)^2 + \ldots]$$

$$= \tau\mu\beta + \frac{\tau\mu(1 - \beta)}{1 - \gamma(1 - \tau)},$$

implying a labor supply of

$$l(w_1) = \mu\beta + \frac{\mu(1 - \beta)}{1 - \gamma(1 - \tau)}. \tag{7}$$

Note that we have derived the equilibrium
the fraction of individuals searching in any
period there are $\tau k(1 - \beta)\gamma$ individuals
time, $\tau k(1 - \beta)\gamma^2(1 - \tau)$ individuals who
time, and so on: the equilibrium unemploy

$$s = \tau(1 - \beta)\gamma[1 + \gamma(1 - \tau) + \gamma$$

or

$$s = \frac{\tau(1 - \beta)\gamma}{1 - \gamma(1 - \tau)}. \tag{8}$$

Since $ds/d\gamma = \tau(1 - \beta)/[1 - \gamma(1 - \tau)]^2$, the equilibrium unemployment rate is an increasing function of γ, as required.

IV. Equilibrium

The discussion above has established conditions that necessarily must hold given the existence of a two-wage dispersion equilibrium. Before using these conditions to investigate the properties of a dispersion equilibrium, we need to examine conditions *sufficient* for existence in terms of the exogenously given parameters of the model in order to ensure that the concept of the dispersion equilibrium is not vacuous.

It is useful to refer back to figure 1 to see what is involved. What needs to be ensured is that some firms have the incentive to "outbid" other, less productive firms.[7] Since we have normalized the support of λ to be $[0, 1]$, this is equivalent to ensuring $0 < \lambda^* < 1$. The cutoff productivity λ^* is easily expressible in terms of the exogenous parameters of the model. From (5) we have

$$\lambda^* = \frac{w_1 l(w_1) - w_0 l(w_0)}{l(w_1) - l(w_0)} = w_1 + \frac{(w_1 - w_0) l(w_0)}{l(w_1) - l(w_0)}$$

$$= w_1 + \frac{(w_1 - w_0)\beta}{(1 - \beta)/[1 - \gamma(1 - \tau)]},$$

using (6) and (7). But $w_1 - w_0 = \tau(v_1 - v_0)/[1 - \gamma(1 - \tau)]$, using (2) and (3); hence

$$\lambda^* = v_1 + b + \frac{\tau(v_1 - v_0)\beta}{1 - \beta}. \tag{9}$$

The condition $0 < \lambda^* < 1$ is thus easily satisfied by a wide range of plausible choices for τ, β, v_0, and v_1.

To see that a "wide range of plausible choices" does indeed exist, it is useful to consider the alternatives. If $\lambda^* \leq 0$, then necessarily $v_1 + b < 0$; that is, even given the existence of unemployment compensation, all workers find the prospect of leisure so loathsome that they would instead be willing to work at a negative wage. In this case the equilibrium outcome will be full employment at the universal (negative) wage of $v_1 + b$. The case of $\lambda^* \geq 1$ comes about if the value placed on leisure by the v_1 individuals and/or the level of unemployment compensation is high enough to induce v_1 individuals to reject all wage offers any firm could profitably offer. In this case (given also that $v_0 + b \leq 1$) the equilibrium outcome will be a single wage of $v_0 + b$ together with an equilibrium unemployment (or nonparticipation) rate of $1 - \beta$.

---

[7] Since unemployment compensation is financed by lump-sum taxation out of dividends (see *n.* 4 above), we need also to ensure that total unemployment compensation payments do not exhaust dividends. That is, on a per capita basis we require that $sb \leq \theta$. For a fixed distribution of productivities this implies an upper limit on b. Alternatively, this upper limit can be changed by changing the assumed distribution of productivities. In the analysis below this condition is always assumed to be satisfied.

V. Comparative Statics

We are now in a position to examine how the equilibrium wage distribution (i.e., w_0, w_1, and γ) varies with the level of unemployment compensation.

Proposition 1: A general increase in unemployment compensation has the following effects on the equilibrium wage distribution:

$$\frac{d\gamma}{db} = \frac{[1 - \gamma(1 - \tau)][a(\lambda^*) - (1 - \gamma)a(w_0)]}{\Delta}, \tag{10a}$$

$$\frac{dw_0}{db} = \frac{[1 - A(w_0)][1 - \gamma(1 - \tau)] - a(\lambda^*)(1 - \tau)(w_1 - w_0)}{\Delta}, \tag{10b}$$

$$\frac{dw_1}{db} = 1, \tag{10c}$$

where $\Delta = [1 - A(w_0)][1 - \gamma(1 - \tau)] - a(w_0)(1 - \gamma)(1 - \tau)(w_1 - w_0)$.

Proof: Rewrite the equilibrium conditions as

$$[1 - \gamma(1 - \tau)]w_0 - (1 - \gamma)(1 - \tau)(w_1 + b) - \tau(w_0 + b) = 0$$
$$A(\lambda^*) - A(w_0) - \gamma[1 - A(w_0)] = 0,$$

regarding γ, w_0, and λ^* as implicit functions of b. Recall from (9) that $d\lambda^*/db = 1$. Differentiating the equilibrium conditions with respect to b then gives the following pair of equations:

$$\begin{bmatrix} (1 - \tau)(w_1 + b - w_0) & [1 - \gamma(1 - \tau)] \\ [1 - A(w_0)] & (1 - \gamma)a(w_0) \end{bmatrix} \begin{bmatrix} d\gamma/db \\ d \end{bmatrix}$$

the solution to which is given by (10a) and
follows directly from (3).

The expressions above are rather formi
that anything is possible depending on the d
tivity index. However, it is possible to derive
results for a broad class of distribution functions.

Proposition 2: Suppose $a'(\lambda^*) \geq 0$, that is, the density function of the productivity index is nondecreasing. Then a general increase in unemployment compensation (i) leads to an increase in the equilibrium rate of unemployment, that is, $d\gamma/db > 0$, and (ii) leads to a "widening" of the wage distribution, that is, $dw_0/db < dw_1/db = 1$.

Proof: (i) If $a'(\lambda) \geq 0$, then $a(\lambda^*) \geq a(w_0)$, implying that the numerator of (10a) is positive. To show that the denominator is unambiguously positive, use

$$A(w_1) = A(w_0) + a(w_0)(w_1 - w_0) + \frac{a'(\bar{w})}{2}(w_1 - w_0)^2$$

for some $\tilde{w}$ between w_0 and w_1. Then rewrite the denominator as

$$\begin{aligned}\Delta &= [1 - \gamma(1 - \tau)][1 - A(w_0) - a(w_0)(w_1 - w_0)] + \tau a(w_0)(w_1 - w_0) \\ &= [1 - \gamma(1 - \tau)][1 - A(w_1) + \frac{a'(\tilde{w})}{2}(w_1 - w_0)^2] \\ &\quad + \tau a(w_0)(w_1 - w_0),\end{aligned}$$

which is positive since $1 - A(w_1) > 0$ and $a'(\tilde{w}) \geqslant 0$. (ii) The result for dw_0/db follows from $a(\lambda^*) > (1 - \gamma)a(w_0)$.

PROPOSITION 3: Suppose again that $a'(\lambda) \geqslant 0$. Then a selective increase in unemployment compensation granted only to those with a low imputed value of leisure leads to a decrease in the equilibrium unemployment rate.

PROOF: The selective increase in unemployment compensation implies

$$\frac{d\lambda^*}{db} = -\frac{\tau\beta}{1 - \beta}.$$

Differentiating the equilibrium conditions with respect to b then yields

$$\begin{bmatrix}(1 - \tau)(v_1 + b - w_0) & 1 - \gamma(1 - \tau) \\ 1 - A(w_0) & a(w_0)(1 - \gamma)\end{bmatrix}\begin{bmatrix}d\gamma/db \\ dw_0/db\end{bmatrix} = \begin{bmatrix}\tau \\ -a(\lambda^*)[\tau\beta/(1 - \beta)]\end{bmatrix}$$

or

$$\frac{d\gamma}{db} = \frac{-\tau a(w_0)(1 - \gamma) - [1 - \gamma(1 - \tau)]a(\lambda^*)[\tau\beta/(1 - \beta)]}{\Delta} < 0,$$

where $\Delta > 0$ is as given in the proof of the second proposition.

The intuition behind these results is not difficult. A general increase in unemployment compensation has the direct effect of increasing the high wage offer since the reservation wage of those who place the highest value on leisure varies directly with b. The effect on the low wage offer is less clear-cut. On the one hand, an increase in unemployment compensation increases the per period utility value of search. On the other hand, if γ increases, the probability that search will pay off decreases. Thus, so long as $d\gamma/db > 0$, the low wage offer will increase by less than the high wage offer. The presumption that γ will increase follows from the necessary increase in λ^*, the cutoff productivity. That is, an increase in unemployment compensation necessarily implies that the *number* of firms offering w_1 must decrease.

The condition that the density function of λ be nondecreasing ensures that the decrease in the number of firms offering w_1 is not more than offset by any decrease in the number of firms offering w_0.

However, if the increase in unemployment compensation is directed solely toward those with a low imputed value of leisure, then w_1 is unaffected. Individuals with the low leisure value begin to search more aggressively and the cutoff productivity λ^* falls. Again, the condition that the density function of λ be nondecreasing ensures that a decrease in λ^* translates to a decrease in γ.

VI. Efficiency

We have established that a general increase in unemployment compensation leads to increased unemployment for a broad class of distribution functions of productivity. One should not, however, be tempted to use this result to conclude that unemployment compensation is inefficient, that is, that the socially optimal level of unemployment compensation is zero.

Suppose the social objective function is per capita utility, $u^* = x^* + v_1 s$, that is, per capita consumption plus the value imputed to per capita leisure. To derive equilibrium per capita production (= equilibrium per capita consumption), first compute total production as the sum of production from low-wage firms and high-wage firms. Total production from firms offering w_0 may be computed as the product of three terms—(i) the number of firms offering
$A(w_0)]$), (ii) $l(w_0)$, and (iii) the average prod
w_0 (= $\int_{w_0}^{\lambda^*}\lambda dA(\lambda)/[A(\lambda^*) - A(w_0)]$). That is,
wage firms is simply $nl(w_0)\int_{w_0}^{\lambda^*}\lambda dA(\lambda)$, and
tion from high-wage firms is $nl(w_1)\int_{\lambda^*}^{1}\lambda dA($
capita consumption is

$$x^* = \frac{n}{k}\left[l(w_0)\int_{w_0}^{\lambda^*}\lambda dA(\lambda) + l(w\right.$$

$$= \frac{1}{\mu[1 - A(w_0)]}\left\{l(w_0)\int_{w_0}^{1}\lambda dA(\lambda)\right.$$

$$\left. + [l(w_1) - l(w_0)]\int_{\lambda^*}^{1}\lambda dA(\lambda)\right\}.$$

or

$$x^* = \beta\int_{w_0}^{1}\frac{\lambda dA(\lambda)}{1 - A(w_0)} + \frac{(1 - \beta)(1 - \gamma)}{1 - \gamma(1 - \tau)}\int_{\lambda^*}^{1}\frac{\lambda dA(\lambda)}{1 - A(\lambda^*)} \quad (11)$$

Let $s_1 = \tau\gamma/[1 - \gamma(1 - \tau)]$, that is, the unemployment rate among the v_1 individuals. Then

$$x^* = \beta \int_{w_0}^{1} \frac{\lambda dA(\lambda)}{1 - A(w_0)} + (1 - \beta)(1 - s_1) \int_{\lambda^*}^{1} \frac{\lambda dA(\lambda)}{1 - A(\lambda^*)},$$

so that per capita utility is given by

$$\begin{aligned} u^* = \beta \int_{w_0}^{1} \frac{\lambda dA(\lambda)}{1 - A(w_0)} &+ (1 - \beta) \int_{\lambda^*}^{1} \frac{\lambda dA(\lambda)}{1 - A(\lambda^*)} \\ &- s \int_{\lambda^*}^{1} \frac{(\lambda - v_1) dA(\lambda)}{1 - A(\lambda^*)}. \end{aligned} \tag{12}$$

The first two terms in (12) give "full-employment output," that is, the hypothetical level of per capita consumption that would be attained with full employment, and the last term gives the utility loss from unemployment.

The temptation to conclude that unemployment compensation must be socially inefficient results from the fact that so long as $ds/db \geq 0$, the utility loss from unemployment is increasing in b. The obvious point is that this temptation founders on the fact that full-employment output also depends on b. An increase in unemployment compensation causes the wage distribution to change in such a way that those who search (the v_1 individuals) are induced to seek out more productive firms. Likewise, if $dw_0/db > 0$, those who do not search will also become employed by more productive firms. The change in the wage distribution drives the least efficient firms out of the market.[8]

The trade-off between these two effects depends on the distribution function of λ. To provide some illustration we examine the simplest possible example, namely, $A(\lambda) = \lambda$, $0 \leq \lambda \leq 1$. The uniform distribution is particularly simple because it allows one to use the equilibrium conditions to find γ (or w_0) as the solution to a simple second-order equation. Once one solves for γ and w_0, the computation of s and u^* is straightforward.

Table 1 presents the equilibrium variables as functions of the level of unemployment compensation for selected values of τ and β, taking $v_0 = 0$ and $v_1 = 0.25$. Starting with $b = 0$, we compute $w_0(b)$, $w_1(b)$, $\gamma(b)$, $s(b)$, and $u^*(b)$ for increments of 0.05 in b up to the point where $\lambda^* \geq 1$. These examples suggest that the optimal level of unemploy-

---

[8] Unemployment compensation thus enhances efficiency by improving the match between workers and firms. This is not, however, a matching model in the standard sense since all individuals are equally productive at any given firm. For a matching model giving an efficiency analysis of unemployment compensation, see Diamond (1981).

TABLE 1

Example Based on Uniform Distribution of Productivities ($v_0 = 0$, $v_1 = .25$. Equilibrium Values of w_0, w_1, γ, s, u^* as Functions of b)

	$b = 0$	$b = 05$	$b = 10$	$b = 15$	$b = 20$	$b = 25$	$b = 30$	$b = 35$	$b = 40$	$b = 45$	$b = 50$	$b = 55$	$b = 60$	$b = 65$	$b = 70$
$\tau = 1, \beta = 1$															
w_0	22	27	32	37	42	47	52	57	62	67	72	77	82	87	90
w_1	25	30	35	40	45	50	55	60	65	.70	75	80	85	90	95
γ	037	040	043	046	050	055	061	069	079	092	110	137	182	269	509
s	003	004	004	004	005	005	006	007	008	009	011	014	.020	032	.084
u^*	624	648	673	698	723	747	772	796	821	844	.868	891	912	927	913
$\tau = 1$ $\beta = 9$															
w_0	21	26	31	36	41	45	50	54	58	61	59				
w_1	25	30	35	40	45	50	55	60	65	70	75				
γ	332	355	382	414	451	496	570	616	701	809	939		$\lambda^* > 1$		
s	005	005	006	007	008	009	011	014	019	030	061				
u^*	618	642	666	[illegible]		735	757	777	793	799	769				
$\tau = 5$ $\beta = 1$															
w_0	11	16	21			36	41	45	50	55	59	63	67	71	73
w_1	25	30				50	55	60	65	70	75	80	85	90	95
γ	170	183				[illegible]	266	298	327	369	422	491	583	706	867
s	083	089				[illegible]	138	155	176	203	241	293	370	491	688
u^*	593	619				[illegible]	701	713	721	725	721	706	670	599	467

ment compensation can be quite "high" relative to wages actually paid,[9] "despite" the fact that $s(b)$ is increasing in b, as indicated by our second proposition.

VII. Conclusion

In this paper we have developed a simple general equilibrium model of search unemployment and used the model to analyze the effects of unemployment compensation. The key feature of the model is the endogeneity of the wage offer distribution. We are thus able to (i) establish the logical consistency of sequential search as a general equilibrium phenomenon and (ii) analyze the comparative statics effects of increases in unemployment compensation taking into account the induced change in the wage offer distribution.

Our results on the effects of increasing unemployment compensation differ significantly from the corresponding effects predicted by the standard "partial-partial" model. Although one *can* expect a general increase in unemployment compensation to increase the equilibrium rate of unemployment, the direct incentive effect for individuals can be offset by firms' adjustment of the wage offer distribution to a considerable degree. More strikingly, one can expect a selective increase in unemployment compensation, made available to those who impute a relatively low value to leisure, to decrease the equilibrium rate of unemployment. This latter result is straightforward in our model but absurd under the standard approach.

Our results on the efficiency aspects of increasing unemployment compensation are also straightforward. The point is that increases in unemployment compensation bring about a reallocation of workers to more productive firms as a result of the change in the wage offer distribution. Of course this result is a direct consequence of the assumption about firm heterogeneity we made to close the model. However, it should be understood that the existence of an equilibrium wage dispersion seems to require assumptions that inevitably imply some sort of reallocation effect.

Although our model is of course stylized in many respects, it should be possible to weaken several of the key assumptions. Some obvious extensions would be (i) to assume more than two classes of individuals, (ii) to allow for a more general (e.g., nonadditive) utility function, and (iii) to consider alternative systems of financing unemployment compensation (e.g., by lump-sum taxation out of profits or by a uniform tax on wages). The main effect of these extensions seems to be to complicate the arithmetic, eliminating the determinate, and thus didactic, flavor of the results. However, the primary moral of this paper, the need to explicitly take into account the endogeneity of the wage offer distribution, undoubtedly emerges unscathed.

---

[9] Of course, this is to some degree an artifact of our choices for v_0 and v_1 in the example

A much more difficult extension, and one that is beyond the scope of this paper, would be to relax the steady-state nature of the analysis to explicitly study the dynamics associated with a change in the level of unemployment compensation. The reason that dynamics would be so difficult in a model of this type is that the reactions of both sides of the market (i.e., both individuals and firms) to a perturbation to an initial steady-state equilibrium need to be modelled, with each agent's optimal reaction depending on the optimal reactions of all other agents. One can reasonably expect, however, that dynamic considerations would not change the basic thrust of the analysis since, as a referee has helpfully emphasized, there is no stock variable in the model that is affected by the path taken from one steady state to another.

Finally, we conclude with an appeal to the empirically minded not to reject equilibrium models of search unemployment as irrelevant theorizing. Although the predictions of the "partial-partial" search model seem to be regarded as firmly established in the empirical folklore, we remain doubtful. First, at least one group of econometricians (Atkinson et al. 1984) has suggested that existing clear-cut empirical results on the effects of unemployment compensation are to some extent an artifact of artful specification. Second, and more fundamental, most empirical work in this area seems
question. The typical cross-section or panel
question of whether individuals who receiv
employment compensation search more tha
generous compensation, holding other diffe
uals constant. But the relevant policy questio
characterized by relatively more generous une
tion have more search unemployment than co
generous compensation. Such a question requires an equilibrium answer.

References

Atkinson, A. B.; Gomulka, J.; Micklewright, J.; and Rau, N. "Unemployment Benefit, Duration, and Incentives in Britain: How Robust Is the Evidence?" *J. Public Econ.* 23 (February/March 1984): 3–26.

Axell, Bo. "Search Market Equilibrium." *Scandinavian J. Econ.* 79, no. 1 (1977): 20–40.

Burdett, Kenneth, and Mortensen, Dale T. "Search, Layoffs, and Labor Market Equilibrium." *J.P.E.* 88 (August 1980): 652–72.

Diamond, Peter A. "Mobility Costs, Frictional Unemployment, and Efficiency." *J.P.E.* 89 (August 1981): 798–812.

Ehrenberg, Ronald G., and Oaxaca, Ronald L. "Unemployment Insurance, Duration of Unemployment, and Subsequent Wage Gain." *A.E.R.* 66 (December 1976): 754–66.

Feldstein, Martin S. "Temporary Layoffs in the Theory of Unemployment." *J.P.E.* 84 (October 1976): 937–57.

Lippman, Steven A., and McCall, John J. "The Economics of Job Search: A Survey: Part I." *Econ. Inquiry* 14 (June 1976): 155–89. (*a*)

———. "The Economics of Job Search: A Survey." *Econ. Inquiry* 14 (September 1976): 347–68. (*b*)

Reinganum, Jennifer F. "A Simple Model of Equilibrium Price Dispersion." *J.P.E.* 87 (August 1979): 851–58.

Rosen, Sherwin. "Unemployment Insurance and Job Search Decisions: Comment." *Industrial and Labor Relations Rev.* 30 (July 1977): 518–20.

Rothschild, Michael. "Models of Market Organization with Imperfect Organization: A Survey." *J.P.E.* 81 (November/December 1973): 1283–1308.

Salop, Steven, and Stiglitz, Joseph E. "Bargains and Ripoffs: A Model of Monopolistically Competitive Price Dispersion." *Rev. Econ. Studies* 44 (October 1977): 493–510.

Topel, Robert, and Welch, Finis. "Unemployment Insurance: Survey and Extensions." *Economica* 47 (August 1980): 351–79.

The Effect of the Discount Rate on Depletion of Exhaustible Resources

Y. Hossein Farzin

World Bank

This paper argues that, despite its widespread acceptance in the literature, the basic proposition that a reduction (an increase) in the discount rate leads to greater conservation (faster depletion) of an exhaustible resource is not generally valid. It shows that the effect of the discount rate on the rate of resource depletion depends on capital requirements for production of the substitute and extraction of the resource and on the size of the resource stock. In fact, it shows that there always exist at least two ranges of the resource stock size for which the effect is opposite to that conventionally believed.

I. Introduction

An increase in the discount rate leads to a fas
ible resources, a decrease to a slower dep
proposition in the economics of exhaustible
tion goes back to the classic work of Hotelling
tant in the current literature, particularly in the
the rates of extraction of exhaustible resources
social point of view. Thus Solow (1974, p. 8), refe [illegible]
(1931) work, writes:

> Hotelling mentions, but rather pooh-poohs, the notion that market rates of interest might exceed the rate at which soci-

---

The views expressed in this paper are mine and should not be interpreted as reflecting the views of the World Bank or any of its affiliated organizations. Valuable comments by an unknown referee, by Geoffrey Heal, and particularly by James Mirrlees at various stages of this paper are gratefully acknowledged. The paper is based on a section of a chapter of my D.Phil. thesis at Oxford University.

[1] For a formal proof of this proposition, see Koopmans (1974).

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0005$01.50

> ety would wish to discount future utilities or consumer surpluses. I think a modern economist would take that possibility more seriously. It is certainly a potentially important question, because the discount rate determines the whole tilt of the equilibrium production schedule. If it is true that the market rate of interest exceeds the social rate of time preference, then scarcity rents and market prices will rise faster than they "ought to" and production will have to fall correspondingly faster along the demand curve. Thus the resource will be exploited too fast and exhausted too soon.

Similarly, Kay and Mirrlees (1975, p. 163) have argued that "a general bias in the economy to consume now and leave too little to our children or our future selves . . . would be reflected in high rates of interest, which will lead to somewhat more rapid depletion of resources."

An explicit statement of the comparative static proposition implicit in the arguments above is provided by Meade (1975, p. 117): "This reduction in the market rate of interest . . . would stimulate investment in oil-in-the-ground to carry oil over from this year's to next year's consumption, or, in other words, it would encourage the conservation in the ground of oil which would otherwise have been pumped for use this year."

My object in this paper is to investigate the validity of the above comparative static proposition, which is without doubt a central one in the economic theory of exhaustible resources. To this end, in Section II I develop a simple resource-substitute model that explicitly allows for the need for capital (as well as other inputs) in both extraction of the resource and production of a substitute. In Section III, I use this model to argue that, despite a widespread belief in it, the conventional wisdom that a reduction (increase) in the rate of interest would lead to a slower (faster) depletion of exhaustible resources is not generally valid. I show that, for exhaustible resources, the allocative effect of a change in the discount rate depends on capital requirements for the development and production of the substitute, on capital costs in resource extraction, and in particular on the size of the stock. Indeed, I show that there always exist at least two ranges of the resource stock size for which the conventional proposition is invalid. Section IV offers some concluding remarks.

II. The Model

Consider a competitive economy that has the following characteristics:[2]

---

[2] The arguments to be made will hold equally for a socially planned economy.

C_1: It is endowed with a known finite stock of an exhaustible resource, S.

C_2: There is an extraction technology that, for the sake of simplicity, is assumed to possess constant returns to scale and to remain unchanged over time. Given this technology and the set of input prices in the economy, the unit cost of resource extraction can be described by the cost function

$$c = c(r, \mathbf{w}), \tag{1}$$

where c is the minimum cost of extracting a unit of the resource at any time, r is the interest rate on capital, and $\mathbf{w}$ is the vector of prices of all other inputs. The cost functions used hereafter are taken to be continuous in their arguments and continuously differentiable.

C_3: The benefits from the use of the resource are measured in monetary units by $V(p_t)$, where

$$V(p_t) \equiv U[x(p_t)] \tag{2}$$

and where p_t denotes the price of a unit of the resource at time t and $x(p_t)$ denotes the market demand function that is assumed to satisfy $x'(p_t) < 0$ and, to avoid corner problems, $\lim_{p_t \to \infty} x(p_t) = 0$ and $\lim_{p_t \to 0} x(p_t) = \infty$. Assuming that the marginal utility of income is constant, the gross benefit (consumer surplus) therefore is evaluated from the market demand function according to $U(x) \equiv \int_0^x p(x)dx$, where $p(x)$ is the inverse of the demand function $x(p)$. So, differentiating (2) with respect to p_t and noting that $U'(x) = p(x)$, one has

$$V'(p_t) = p_t x'(p_t).$$

I shall presently introduce a perfect substit that (2) and (3) would be interpreted as refer and its substitute.

C_4: Future benefits are discounted at a co assumption that capital markets are perfect. rate of interest.

C_5: Following Nordhaus (1973), I assume that the technology [illegible] economy allows the procurement of a perfect substitute for the exhaustible resource, possibly at a cost that is higher than the current price of the resource, but on an effectively inexhaustible resource base. For instance, we can think of crude oil as the exhaustible resource and the breeder reactor as an approximation of the substitute technology. Assume that the economy already has access to this technology and that there is no capacity constraint on production of the substitute. For expository convenience, also assume that returns to scale in production of the substitute are constant, so that the unit production cost of the substitute can be written as

$$p = p(r, \mathbf{w}), \tag{4}$$

where p is the minimum production cost per unit of the substitute, which, for any given factor price, exceeds the unit extraction cost of the resource, c.[3]

The problem facing such an economy is to choose the price path of the resource p_t and the date T at which the substitute should replace the resource to maximize W, the present net benefit from the depletion of the resource stock, and the use of the substitute. Formally, the problem is[4]

$$\max_{\{p_t\},T} W = \int_0^T e^{-rt}[V(p_t) - x(p_t)c(r, \mathbf{w})]dt + \int_T^\infty e^{-rt}[V(p_t) - x(p_t)p(r, \mathbf{w})]dt, \tag{5}$$

subject to the constraints $\int_0^T x(p_t)dt \leq S$, $x(p_t) \geq 0$, and $p_t \geq 0$ for all $t \geq 0$.

Forming the Lagrangian associated with problem (5), differentiating it with respect to p_t and T, and using (3), we obtain the following necessary conditions for an optimum:

$$p_t - c(r, \mathbf{w}) = \lambda e^{rt}, \quad t \in [0, T], \tag{5a}$$

$$p_t = p(r, \mathbf{w}), \quad t \in [T, \infty], \tag{5b}$$

and

$$\int_0^T x(p_t)dt = S, \tag{5c}$$

where $\lambda = p_0 - c(r, \mathbf{w})$ is the Lagrangean multiplier and measures the initial scarcity rent (price minus extraction cost) on the resource.

The optimal policy governed by conditions (5a)–(5c) is familiar in the literature (see, e.g., Dasgupta and Stiglitz 1976). It requires that the scarcity rent on the exhaustible resource rise at the rate of discount until the resource price has risen sufficiently to cover the marginal production cost of the substitute, at which time the stock of the resource is completely depleted and the substitute will be supplied at a price equal to its marginal cost of production $p(r, \mathbf{w})$ and at a constant

---

[3] Clearly, if $p < c$ the resource stock will never be used, and if $p = c$ the resource and substitute can be considered to be homogeneous and thus treated as a single good.

[4] In setting problem (5), I have assumed that the economy will use up the stock of the resource before it switches to the substitute. That this is indeed the optimum order in which the resource and substitute should be used is intuitively clear: given the assumption of constant returns in production of the substitute and its higher cost of production (relative to extraction costs of the resource), it will always pay to delay the use of the substitute and consume instead the cheaper resource until its stock is exhausted. For a formal analysis of the optimum order of resource/substitute utilization under various conditions of returns to scale in substitute production, see Farzin (1981).

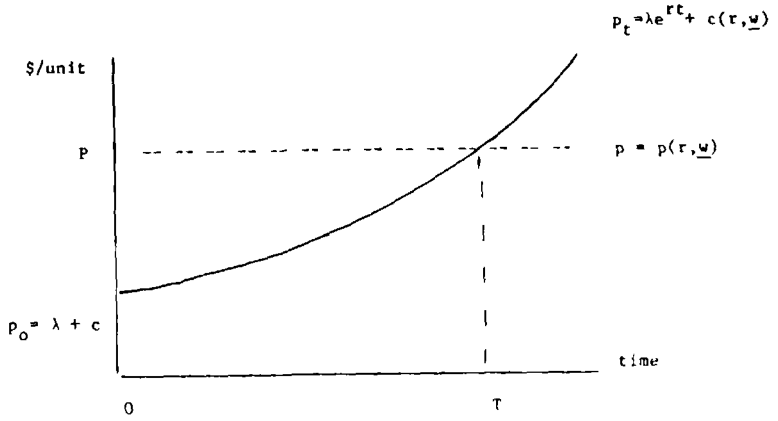


FIG. 1

flow rate of $x[p(r, \mathbf{w})]$ units.[5] The optimal policy is illustrated in figure 1. Further insight into the properties of the optimal policy may be gained from the following comparative static results that can be readily derived from the arbitrage conditions (5a)–(5c) and verified by figure 1:

$$\frac{\partial T}{\partial \bar{p}} > 0, \quad \frac{\partial p_0}{\partial \bar{p}} > 0, \quad \frac{\partial \lambda}{\partial \bar{p}} > 0 \tag{6a}$$

$$\frac{\partial T}{\partial \bar{c}} > 0, \quad \frac{\partial p_0}{\partial \bar{c}} > 0, \quad \frac{\partial \lambda}{\partial \bar{c}} < 0$$

$$\frac{\partial T}{\partial S} > 0, \quad \frac{\partial p_0}{\partial S} < 0, \quad \frac{\partial \lambda}{\partial S}$$

where p_0 is the optimal initial price of the re
respectively, the unit costs of the substitute an
input prices $(r, \mathbf{w})$.

According to (6a) an exogenous increase in the unit cost of the substitute delays the date at which the economy switches from the resource to the substitute, slows down the rate of resource exploitation, and benefits the resource owners by enabling them to collect larger amounts of royalties at any time. A rise in the unit cost of extraction has the same qualitative effects except that it reduces the rents accruing to resource owners. Obviously, these effects will be reversed for a reduction in the unit costs; thus, for example, a cost-reducing technological breakthrough in the substitute and/or the resource sector speeds up the rate of resource extraction and advances the date of utilization of the substitute.

---

[5] That the optimal path should be continuous at $t = T$ must be obvious, for otherwise either the resource will be undercut by the substitute or it will pay to hold back some of the resource stock, no matter how small, until time T when capital gains are huge

The interpretation of inequalities (6c) is straightforward. An addition to the existing reserves of the resource reduces the scarcity rent and hence the price of the resource. In the limiting case where the size of the resource stock becomes infinitely large, the scarcity rent tends to zero, thus implying that the price of the resource in that case will be equal to its marginal cost of extraction. This is akin to economic intuition, for as the stock of the resource becomes very large the resource becomes like a conventional inexhaustible commodity, in which case the efficiency considerations dictate that the resource be priced at its marginal extraction cost. Since this statement proves to be essential in the analysis of the next section, a formal proof of it is presented here.

Using (5a) and (5b), one can write

$$p(r, \mathbf{w}) - c(r, \mathbf{w}) = \lambda e^{rT}. \tag{7}$$

Differentiating (7) and (5c) with respect to S and solving for $\partial\lambda/\partial S$ and $\partial T/\partial S$ gives

$$\frac{\partial\lambda}{\partial S} = -\frac{r\lambda}{x(p_0)} \tag{8}$$

and

$$\frac{\partial T}{\partial S} = \frac{1}{x(p_0)}. \tag{9}$$

Thus, from (8), λ is a decreasing function of S. Furthermore, the limit of λ, as $S \rightarrow \infty$, cannot be positive, for suppose $\lim_{S\rightarrow\infty} \lambda = \hat{\lambda} > 0$. Then, from (7), one must also have $\lim_{S\rightarrow\infty} T = \hat{T} > 0$, in which case the integral $\int_0^T x(c + \lambda e^{rt})dt = S$ obtains a finite limit as $S \rightarrow \infty$, which is a contradiction. It therefore follows that $\lim_{S\rightarrow\infty} \lambda = 0$, which in turn implies that

$$\lim_{S\rightarrow\infty} p_0 = c. \tag{10}$$

III. The Discount Rate Effect

How does a change in the discount rate affect the optimal rate of resource exploitation? Bearing in mind that both unit costs p and c depend on the interest rate r, one can differentiate (5c) and (7) with respect to r to solve for $\partial\lambda/\partial r$ and then use $p_0 = \lambda + c$ to obtain, after some manipulation,

$$\frac{\partial p_0}{\partial r} = \frac{\dfrac{\lambda x(p)}{p - c}\dfrac{\partial c}{\partial r} + r\dfrac{\partial c}{\partial r}\displaystyle\int_0^T e^{-rt}x(p_t)dt - \lambda S}{x(p_0)}, \qquad (11)$$

where it is recalled from the theory of cost functions (see, e.g., Shephard 1970) that $\partial p/\partial r$ and $\partial c/\partial r$ indicate the capital requirement per unit (or capital intensity) of the substitute and resource, respectively.

It is thus evident from (11) that the capital intensities in the two sectors play an important part in determining the effect of a change in the rate of discount on the intertemporal allocation of the resource and that their effects are such that a reduction in the rate of discount could result in a faster extraction of resources (disinvestment in resources in the ground). This runs counter to the conventional proposition that a lower discount rate unambiguously leads to conservation of exhaustible resources. The intuitive belief in the proposition might have been derived from two unrealistic assumptions: first, that there exists no potential substitute for the resource, which leads to the exclusion of the term $\partial p/\partial r$ from (11); and second, that there are no capital costs whatsoever involved in the extraction of the resource, which has the effect of eliminating the term $\partial c/\partial r$ from (11).[6] Alternatively, it might have been derived from the unguided intuition that the act of keeping resources in the ground is an inves
identical to any other form of investment, so t
"reduction in the market rate of interest . .
ment in oil-in-the-ground" (Meade 1975, p
verified by equation (11), a reduction in the
about two counteracting effects. First, inasmu
reflects the rate of time preference, a reductio
use of resources to the future (a conservation
the extent that it reflects the cost of capital services, it
costs in both the substitute and resource sectors and hence induces a faster rate of resource depletion (a disinvestment effect). The net effect depends, among other things, on the degree of capital intensities in the substitute and resource sector, on the cost differential

---

[6] It is interesting to note that when the need for the input of capital services in resource extraction and substitute production is ignored, so that $\partial c/\partial r = \partial p/\partial r = 0$, eq (11) reduces to $\partial p_0/\partial r = -\lambda S/x(p_0) < 0$, which is the conventional comparative static result associated with the simple Hotelling model. Moreover, substituting for $-\lambda S/x(p_0)$ from (8) and noting that $\partial\lambda/\partial S = \partial p_0/\partial S$, one obtains $(r/p_0)(\partial p_0/\partial r) = (S/p_0)(\partial p_0/\partial S)$, i.e., elasticity of p_0 with respect to S, the stock size, must be equal to the elasticity of p_0 with respect to r, the interest rate. This is an empirically testable implication of the Hotelling model that has not been noted in the literature

between the substitute and the resource, and on the size of the resource stock.

One can take the analysis further and use (11) to derive a general condition that unambiguously determines the qualitative role of the discount rate in the allocation of exhaustible resources. Toward this end, let

$$K_R \equiv \int_0^T e^{-rt}\left[rx(p_t)\frac{\partial c}{\partial r}\right]dt \tag{12a}$$

and

$$K_S \equiv \frac{\lambda x(p)}{p - c}\,\frac{\partial p}{\partial r} = \int_T^{\infty} e^{-rt}\left[rx(p)\frac{\partial p}{\partial r}\right]dt, \tag{12b}$$

where K_R and K_S measure, respectively, the present values of capital requirements in the resource and substitute sectors. Also, substituting from (5a) and (5b) for p_t in (5) to obtain the maximized value of W (denoted by W^*), differentiating W^* with respect to S, and making use of (3), (8), and (9), one has

$$\frac{\partial W^*}{\partial S} = \lambda, \tag{12c}$$

where λ is the present shadow price of a unit of resource stock in the ground, so that λS is the present value of the resource stock.

Using (12a)–(12c), we can then rewrite (11) as $\partial p_0/\partial r = [(K_R + K_S) - \lambda S]/x(p_0)$, which implies that $\partial p_0/\partial r \gtreqless 0$ as $(K_R + K_S) \gtreqless \lambda S$. We can therefore state the following proposition.

PROPOSITION 1: A reduction in the discount rate leads to a more rapid depletion of resources if the sum of the present values of capital requirements in the substitute and resource sectors exceeds the present value of the resource stock.[7]

---

[7] It is important to note that in interpreting $r(\partial c/\partial r)$ and $r(\partial p/\partial r)$ as capital costs per unit of the resource and of the substitute, we are making the simplifying assumption that the unit cost, and hence price, of capital goods is fixed and normalized to unity, so that the cost of capital goods services is in effect equal to the interest rate. However, in general, the rental cost of capital goods services will be $R = rp_k$, where p_k, the price (unit cost) of capital goods, depends in turn on the rental cost via $p_k = g(R, \mathbf{w})$. Thus, to be precise, the interest rate arguments of the unit cost functions for the resource and substitute should be replaced by the rental cost, R, to yield $c(R, \mathbf{w})$ and $p(R, \mathbf{w})$, respectively. Adopting this modification and differentiating the expression $R = rg(R, \mathbf{w})$ implicitly with respect to r to obtain $\partial R/\partial r = g[1 - r(\partial g/\partial R)]^{-1}$, it can then be readily verified that $r(\partial c/\partial r) = R(\partial c/\partial R)/(1 - s)$ and $r(\partial p/\partial r) = R(\partial p/\partial R)/(1 - s)$, where $s = d \ln g/d \ln R$ is the cost share of capital services in the capital goods sector. Therefore, the feedback effect of a change in the interest rate on costs of extraction and substitute production is even greater, by a factor of $1/(1 - s)$, than that noted here.

This is a basic result, which has not been discussed in the literature, according to which a faster depletion of resources is always caused by a rise in the discount rate, never by a reduction.

It is clear from (11) that the condition in proposition 1 will be satisfied if the cost differential ($p - c$) is sufficiently small. It is also obvious from (11) that the condition will hold if the capital intensities of the substitute and/or the resource are sufficiently large. What is less obvious, however, is whether the condition will hold when the resource stock is large or when it is small. This is an important question because if it turns out that $\partial p_0/\partial r > 0$ when the stock of the resource is small, then reducing the discount rate in the conventional belief that it will lead to greater resource conservation will have adverse effects that can be far more serious than if $\partial p_0/\partial r > 0$ holds when the resource stock is large. The following analysis therefore examines the sizes of the resource stock for which the condition $\partial p_0/\partial r > 0$ holds.

From the resource stock constraint (5c), note that $\lim_{S \to 0} T = 0$, so that by (7) and the fact that $p_0 = \lambda + c$,

$$\lim_{S \to 0} p_0 = p. \tag{13a}$$

Also recall from (10) that

$$\lim_{S \to \infty} p_0 = c \tag{13b}$$

Since $\partial p_0/\partial r$ is a continuous function of S (which follo
one can interchange the operations of differen
r), taking the limit (on S) to write

$$\lim_{S \to 0}\left(\frac{\partial p_0}{\partial r}\right) = \frac{\partial}{\partial r}\left(\lim_{S \to 0} p\right.$$

and

$$\lim_{S \to \infty}\left(\frac{\partial p_0}{\partial r}\right) = \frac{\partial}{\partial r}\left(\lim_{S \to \infty} p_0\right). \tag{14b}$$

Finally, substituting from (13a) and (13b), respectively, into (14a) and (14b), and remembering that both p and c are increasing in r, we have

$$\lim_{S \to 0} \frac{\partial p_0}{\partial r} = \frac{\partial p}{\partial r} > 0$$

and

$$\lim_{S \to \infty} \frac{\partial p_0}{\partial r} = \frac{\partial c}{\partial r} > 0.$$

We can thus state the following proposition, which is also illustrated in figure 2.

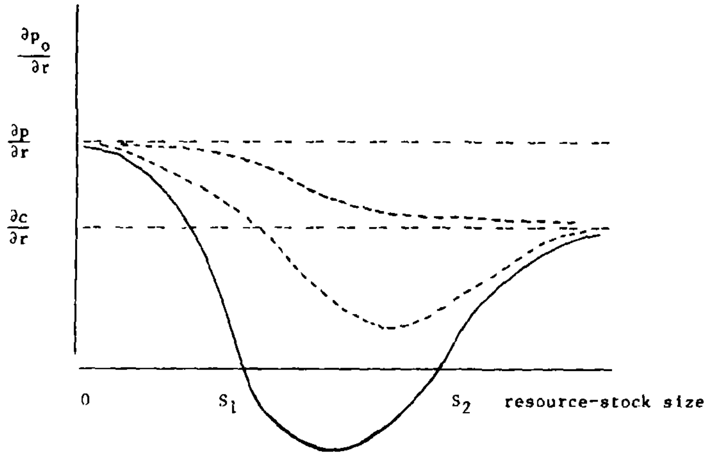


FIG. 2

PROPOSITION 2: There are always at least two ranges of resource stock size for which, contrary to conventional belief, a lower discount rate will lead to a faster depletion of an exhaustible resource and a higher discount rate to a slower depletion. The ranges are when the resource stock is sufficiently small ($S < S_1$ in fig. 2) or when it is sufficiently large ($S > S_2$).

The economic intuition behind this result is simple. As was pointed out earlier, when the stock of the resource is very large, the resource is much like an ordinary product for which the price is determined by the marginal cost of production; as such, a reduction in the rate of interest renders the resource cheaper and hence increases its rate of use. In the opposite case, when the stock of the resource is very small, the resource can enjoy a scarcity rent almost as large as the difference between the cost of producing the substitute and its own extraction cost, implying that it will command a price roughly equal to the production cost of the substitute. In that case, a reduction in the interest rate reduces the cost of the substitute, and hence the price obtainable by the resource, leading to faster use of the resource.

As the dotted curves in figure 2 illustrate, the reversal effect of a change in the discount rate ($\partial p_0/\partial r > 0$) may prevail at all levels of resource endowment. This can occur when the share of capital cost in unit costs of resource extraction and substitute production—that is, when $(p/r)(\partial p/\partial r)$ and $(c/r)(\partial c/\partial r)$—is rather large and/or when the cost differential between the resource and the substitute is small. However, the implication of the reversal effect is particularly important for economies with poor natural resource endowments. For such economies, a lower discount rate, instead of promoting resource conservation, will accelerate resource exhaustion.

IV. Concluding Remarks

In this paper I have argued that, despite its widespread acceptance in the literature, the proposition that a decrease in the discount rate leads to a slower depletion of exhaustible resources, and vice versa, is not generally valid. I have shown that a decrease in the discount rate has two distinct and countervailing effects. The first is the straightforward "conservation effect" of a lower discount rate—consumption in the future is rendered more attractive. The second is the less straightforward "disinvestment effect"—the lower discount rate reduces the unit costs of extraction of the resource and production of the substitute and hence encourages faster rates of resource extraction. This latter effect has been completely neglected in the literature, and yet my analysis has shown that for at least two ranges of resource stock size it is always this effect that dominates. Of course, to maintain the simplicity of the model, I have abstracted from resource exploration activities that no doubt play an important role in making decisions about the rates of resource depletion. However, to the extent that a lower discount rate encourages investment in exploration activities and hence increases long-run resource availability, the inclusion of such activities in the model would reinforce the result of my analysis.

References

Dasgupta, Partha S., and Stiglitz, Joseph F. "Unc traction under Alternative Institutional Arrange Report. Stanford, Calif.: Stanford Univ., 1976

Farzin, Y. Hossein. "Competition and Substitutes haustible Resource." D.Phil. thesis, Oxford Univ

Hotelling, Harold. "The Economics of Exhaustible Res (April 1931). 137–75.

Kay, John A., and Mirrlees, James A. "The Desirability of Natural Resource Depletion." In *The Economics of Natural Resource Depletion*, edited by David W. Pearce and James Rose. London: Macmillan, 1975

Koopmans, Tjalling C. "Proof for a Case Where Discounting Advances the Doomsday." *Rev. Econ. Studies* (symp., 1974), pp. 117–20

Meade, James E. *The Intelligent Radical's Guide to Economic Policy: The Mixed Economy*. London: Allen & Unwin, 1975.

Nordhaus, William D. "The Allocation of Energy Resources." *Brookings Papers Econ. Activity*, no. 3 (1973), pp. 529–70.

Shephard, Ronald W. *Theory of Cost and Production Functions*. Princeton, N.J.: Princeton Univ. Press, 1970.

Solow, Robert M. "The Economics of Resources or the Resources of Economics." *A.E.R. Papers and Proc.* 64 (May 1974): 1–14

An Estimable Dynamic Stochastic Model of Fertility and Child Mortality

---

Kenneth I. Wolpin

Ohio State University and Center for Human Resource Research

This paper develops a finite-horizon dynamic stochastic model of discrete choice with respect to life-cycle fertility within an environment where infant survival is uncertain. The model yields implications for the number, timing, and spacing of children. A tractable estimation method is developed for the linear constraint–quadratic utility case that is intimately tied to the dynamic optimization problem, and the method is applied to Malaysian household data. Estimation is based on integrating the numerical solution of the dynamic programming model of behavior with a maximum likelihood procedure.

Introduction

Connections between child mortality and fertility are at the root of many explanations of the demographic transition and are important for population policy in less developed countries. In this paper, I present an estimable dynamic stochastic model of life-cycle fertility within an environment where infant survival is uncertain. The optimization model provides a unified treatment of replacement behavior (Ben-Porath 1976; Schultz 1976; Olsen 1980), that is, the reaction to a realized infant death; of generalized survival uncertainty, that is, the

I am grateful for research support received from the William and Flora Hewlett Foundation and from NSF grant no. SES-8207594. Zvi Eckstein contributed many useful comments and substantially enhanced my understanding and appreciation of dynamic economic models. Randall Olsen provided extremely valuable comments, particularly, though not exclusively, with respect to the estimation methodology. Comments from Jim Jordan, T. Paul Schultz, Paul McGuire, and the editor are gratefully acknowledged. Paul McGuire most ably performed the rather complicated calculations.

[*Journal of Political Economy*, 1984, vol. 92, no. 5]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0004$01.50

reaction to the mortality environment; and of spacing behavior. The model yields implications for the number, timing, and spacing of children in such an environment.[1] A tractable estimation method is developed for the linear (constraint)–quadratic (utility) case that is intimately tied to the dynamic optimization problem, and the method is applied to data from Malaysia. Although the econometric methodology is interesting in its own right, the paper also provides empirical estimates of important substantive phenomena in the context of a sequential fertility model.[2]

The finite-horizon dynamic programming problem that is assumed to be solved by the individual yields time-variant decision rules concerning the choice whether or not to have a child in each period. Analytical representation of those decision rules is not feasible. Estimation is based on integrating the numerical solution of the dynamic programming model with a maximum likelihood procedure. The complexity of the behavioral model is limited by the computational burden of the dynamic programming solution algorithm given the iterative nature of the estimation procedure. The methodology, however, is general to any finite-horizon dynamic stochastic model that can be numerically solved and would be applicable, for example, to job search models, labor supply models, and schooling choice models.[3]

The methodological approach is highly structured and follows the tradition of Heckman and MaCurdy (1980) in the analysis of life-cycle labor supply or of Sargent (1978) in macro time-series analysis. Because of the discipline this approach imposes on
generally the case (and there is no excep
quite simple, even if analytical solutions a
tion.[4] This characteristic has led to fund
approach, not necessarily as a research a
mediate tool for addressing substantive is-

---

[1] Spacing involves a fertility pattern in which births are not all consecutive in the context of discrete time. Timing relates to life-cycle stage. Births may occur early or late in the childbearing period, and they may be closely or widely spaced. The model presented here explicitly characterizes the birth sequence and thus solves for both aspects of fertility.

[2] Sequential fertility models have been studied by Heckman and Willis (1976), Hotz (1980), Moffitt (1980), and Vijverberg (1981). The model presented here is very similar to that of Heckman and Willis. None of these studies develop the econometric methodology necessary to provide estimates of "structural" parameters.

[3] Other very recent examples of the type of estimation method developed here are Miller's (1982) job-matching model and Pakes's (1983) patent model. Issues related to the estimation of dynamic programming models in job search contexts can be found in Flinn and Heckman (1982). Work by Hansen and Sargent (1981) is related.

[4] In a naive sense, all models are estimable save for computational expense. See Hartley (1981) who advocates a much greater effort in this direction.

The advantages and disadvantages of structural estimation as opposed to econometric specifications as "approximations" to optimizing models are well known. A practical issue arises, however, if the approaches lead to different substantive results but fit the data equally well. The structural model, if correctly specified, provides legitimate restrictions that permit more precise inference and a more parsimonious representation of complex relationships; but if the model is incorrectly specified, all inferences may be contaminated regardless of the offending assumptions. Relationships estimated as "approximations to theory," since they do not explicitly follow from theory, contain implicit and unknown assumptions but are in a sense less restrictive and may provide more accurate, though cruder, estimates of particular responses of interest. The choice of estimates is likely to be problem specific. Those who have strong reservations about the substantive implications of the structural model estimated here can view this as a prototype of potentially more fruitful models yet to be developed.

The paper is organized as follows. In the next section I specify a *discrete-time, discrete-outcome dynamic stochastic fertility model* with exogenous but uncertain infant mortality and income. I then describe the estimation strategy and present the data and results. A discussion section interprets the results and briefly addresses some of the broader methodological issues. The final section summarizes.

A Discrete-Time, Discrete-Outcome Dynamic Stochastic Model of Fertility

The household is assumed to possess an intertemporally separable utility function defined over a finite horizon having as arguments the number of surviving children and a single composite consumption good. There is no direct consumption value to the spacing or timing of births except insofar as the future is discounted. A period is defined to be that length of time within which a single birth may occur.[5] Births may occur only during the fertile stage, the length of which is known with certainty to be T periods.[6] There are $\tau - T$ subsequent periods of sterility. Contraception is assumed to be perfect during the fertile stage, and, conversely, conception can be timed without error. The choice is of a discrete nature, that is, to have a child or not.[7] There is a fixed cost of bearing a live child whether or not the child survives the first period. There is also a technologically determined fixed cost of maintaining a live child for the first period of its life.[8] Infant mortality and household income are exogenous, and each follows a stochastic process known to the household.[9]

---

[5] As in all discrete-time models the length of a period is somewhat arbitrary. In the case of fertility, however, biological considerations provide some guidance

[6] The model would be significantly more complicated to estimate if T were itself a random variable

The decision problem for household i at time t is given by

$$\max E_t \sum_{k=0}^{\tau-t} \delta^k U^i_{t+k}(M^i_{t+k}, X^i_{t+k}), \tag{1}$$

with respect to n^i_t, where $n^i_t = 1$ indicates a birth at t and $n^i_t = 0$ indicates no birth; X^i_{t+k} is the level of goods consumption in period $t + k$ and M^i_{t+k} is the stock of surviving children. The utility function is assumed to be quadratic, namely, at any period t:

$$U^i_t(M^i_t, X^i_t) = (\alpha_1 + \xi^i_t)M^i_t - \alpha_2 M^{i^2}_t + \beta_1 X^i_t - \beta_2 X^{i^2}_t + \gamma M^i_t X^i_t. \tag{2}$$

The taste parameter ξ^i_t is assumed to be random with zero mean and finite variance.[10] The law of motion for M_t (without the i superscript) is

$$\begin{aligned} M_t &= M_{t-1} + n_t - d_t, & t &= 1, \ldots, T, \\ M_t &= M_T, & t &= T + 1, \ldots, \tau, \end{aligned} \tag{3}$$

with initial condition

$$M_0 = 0. \tag{4}$$

The variable d_t is unity if a death occurs during p
born at t and is zero otherwise. A child is assu
that is, longer than the parents, if it surv
equation (1), E_t is the expectations operato

---

[7] Introducing contraceptive risk would change the
The choice would become whether or not to contrac
traceptive behavior or on births could be used in estimati
on infant survival, accommodating additional forms of uncert
bitious.

[8] It is, of course, more natural to assume that the maintenance cost of a child is positive at older ages, as has been assumed in other theoretical life cycle fertility models (e.g., Hotz 1980). However, doing so expands the dimensionality of the state space from one that is equal to the number of surviving children in any particular period (i.e., T possible states in period T) to one in which the size depends on potential sequences of children. If the maintenance cost changed each period of a child's life, e.g., there would be 2^T possible state values to consider at period T.

[9] I ignore the possibility that parental expenditures can influence infant mortality in order to avoid modeling a second decision variable.

[10] This particular placement of ξ^i_t in the utility function is purposeful as will become apparent when the estimation method is presented.

[11] In the application below and in many other LDC contexts a substantial component of child mortality occurs within the first few years of life.

mation set as of t and δ is the discount factor. Period 1 is the first period in which a woman makes an independent decision about fertility.[12]

Child deaths are generated as follows:

$$\begin{aligned} d_t &= 0 \text{ iff } \mathbf{G}_t\pi_1 + u_t \geqslant 0 \\ d_t &= 1 \text{ iff } \mathbf{G}_t\pi_1 + u_t < 0, \end{aligned} \tag{5}$$

where the $\mathbf{G}_t$ vector may contain time-varying variables, for example, calendar time, or possibly household characteristics. The random component u_t has zero mean and finite variance.

The budget constraint is assumed to be satisfied period by period:

$$Y_t = X_t + b(n_t - d_t) + cn_t, \tag{6}$$

with Y_t the household income at t, c the fixed cost of a birth, and b the maintenance cost of a surviving child. Income at time t has exogenous determinants H_t, for example, age, and is given by

$$Y_t = \mathbf{H}_t\pi_2 + v_t, \tag{7}$$

where v_t has zero mean and finite variance.

It is important to specify precisely the content of the information set. As of period t the household is assumed to know the number of surviving children up to period t, M_{t-1}. The systematic components of income and infant mortality are also assumed known at each t for all $t = 1, \ldots, \tau$, that is, $\mathbf{G}_1\gamma_1, \ldots, \mathbf{G}_t\pi_1, \mathbf{H}_1\pi_2, \ldots, \mathbf{H}_\tau\pi_2$ is known at the initial period. Past realizations of income are obviously known, but neither the current realization (v_t) nor future realizations ($v_{t+1}, \ldots, v_\tau$) are assumed known.[13] Similarly, past deaths are known, but obviously not the fate of the current (to be born) child or future children. The household is also assumed to know its current (and past) tastes, ξ'_t, though not those it will have in the future. Finally, the joint distribution (conditional at t) of the random income, mortality, and taste components is assumed known for all $t = 1, \ldots, \tau$ at the initial period.

The maximization problem given by (1)–(7) can be solved using the principle of optimality (Bellman 1957). There is only a single (discrete) control variable (n_t) and a single state variable (M_{t-1}). The solution is obtained by backward recursion. Decision rules, which entail comparisons between expected utility levels with and without a birth, are not time invariant given the finite decision horizon (T) and are difficult to represent analytically.

---

[12] Period 1 does not necessarily correspond to the date of marriage, since the decision not to marry may be a childbearing decision.

[13] Thus, the fertility decision at t is assumed to be made prior to the realization of income, which captures the fact that the decision to have a child precedes the birth by the gestation lag.

Some features of the model can be explored analytically within a two (decision) period framework though in a longer-horizon model they can most efficiently be exhibited through simulation.[14] The most important features are these:[15] (1) It is capable of generating a wide variety of birth sequences, that is, timing and spacing patterns. Given the assumption that the household cannot save or dissave, a rising income profile may generate spaces (and delay) as it counteracts the desire to have children as fast as possible.[16] (2) The model is capable of generating a wide variety of replacement patterns from full replacement of any child death to zero replacement. Replacement patterns and spacing are highly interdependent phenomena. (3) The model is not likely to generate higher fertility in high mortality environments, that is, where the infant survival probability is low. Essentially, since a child can die only in its first period, it will be optimal to follow a sequential replacement strategy unless children are desired at the end of the fertile stage where replacement is physiologically not possible.

Method of Estimation

To focus on the central estimation problem, assume that the income and survival rate functions (eqq. [5] and [7]) are known to the researcher.[17] Also, for ease of presentation, assume that ξ_t^i, u_t^i, and v_t^i are identically independently distributed in both the household (i) and period (t) dimensions and are stochastically independent. These assumptions can all be relaxed in principle to permit generalized vector time-series representations in the sense that the est
outlined below would still be valid. Howeve
sentation that follows would need to be su

Now, with these assumptions and given,

---

[14] See Wolpin (1982) for a more complete discussi
and for a presentation of simulation results.

[15] The simulation results assume that the individual has
for all t.

[16] Heckman and Willis (1976) present a model that differs in the existence of contraceptive uncertainty rather than survival uncertainty. The spacing-timing result is conjectured there. Spacing could also arise if capital markets were perfect and the rate of interest exceeded the rate of time preference. That modeling strategy was not pursued since any full-solution estimation method would have required solving for optimal savings behavior along with optimal fertility behavior. Neither extreme savings assumption would seem appropriate in the Malaysian context.

[17] If they are estimated by the researcher, then efficient estimation of the entire structural model would require joint estimation of these relationships with the fertility decision rules. Procedurally, this is a straightforward task, though in the empirical application it is ignored for computational reasons. It would also permit a test of cross-equation restrictions and, in this sense, of the model.

form of the utility function, the decision to have a child at any arbitrary period t is determined by the sign of the following expression:

$$J_t = E_t(LU_t|n_t = 1, \xi_t = 0) - E_t(LU_t|n_t = 0, \xi_t = 0) + P_t\xi_t, \quad (8)$$

where LU_t is lifetime utility at t and P_t is the infant survival probability at t.[18] The decision rule given by (8) is deterministic for the household; that is, the household decides to have a child if J_t is positive and decides not to have a child if it is not. If the researcher does not know ξ_t, then from that perspective the fertility decision is random. Assuming that the conditional expectations expressions in equation (8) can be calculated in a numerical sense for given parameter values by the researcher, the estimation procedure is straightforward. Obviously, there exists a unique value of ξ_t for which J_t is exactly zero.[19] The critical value, say ξ_t^*, depends on the value of the parameters that are to be estimated as well as the data on income and mortality.[20]

If ξ_t is assumed to follow a normal distribution, with variance σ^2,

---

[18] Recall that $E_t\xi_t = \xi_t$; i.e., the household knows its draw prior to making its decision.

[19] The placement of the estimation error in α_1 was governed entirely by computational considerations. Any other parameters of the model could have been assumed random, α_2, β_2, γ, b, c, or δ. Furthermore, an estimation error could have been introduced into the income or survival probability functions, i.e., as variables unknown to the researcher but known to the household. Some of these other choices would have destroyed the simple form of eq. (8). There would be no guarantee of a unique critical value and all solutions would have to be located numerically.

[20] As an example, the decision rule at T is given by

$$\begin{aligned} J_T = P_T\Bigg[&\alpha_1\Big(1 + \sum_{t=T+1}^{T} \delta^{t-1}\Big) - \alpha_2(2M_{T-1} + 1)\Big(1 + \sum_{t=T+1}^{T} \delta^{t-1}\Big) \\ &- \beta_2(b^2 - 2bH_T\pi_2 + 2cb) \\ &+ \gamma\Big(-bM_{T-1} + H_T\pi_2 - c - b + \sum_{t=T+1}^{T} \delta^{t-1}H_t\pi_2\Big)\Bigg] \\ &+ c[-\beta_1 - \beta_2(c - H_T\pi_2 - \gamma M_{T-1})] \\ &+ P_T\xi_T \end{aligned}$$

The following conditional expectations are necessary for the derivation:

$$E_{t-j}(X_t|n_t = 1, d_t = 0) = H_t\pi_2 - c - b,$$

$$E_{t-j}(X_t|n_t = 1, d_t = 1) = H_t\pi_2 - c,$$

$$E_{t-j}(X_t|n_t = 0, d_t = 0) = H_t\pi_2,$$

$$E_{t-j}(X_t^2|n_t = 1, d_t = 0) = (H_t\pi_2)^2 + \sigma_v^2 + c^2 + b^2 - 2(c + b)H_t\pi_2 + 2cb,$$

$$E_{t-j}(X_t^2|n_t = 1, d_t = 1) = (H_t\pi_2)^2 + \sigma_v^2 + c^2 - 2cH_t\pi_2,$$

$$E_{t-j}(X_t^2|n_t = 0, d_t = 0) = (H_t\pi_2)^2 + \sigma_v^2,$$

for $j = 0, \ldots, t$. It should be noted that although $(H_t\pi_2)^2$ and σ_v^2 appear in expressions for $E_T(LU_T|n_T = 1)$ and $E_T(LU_T|n_T = 0)$, they do not appear in J_T. In the 1-period model, certainty equivalence obtains.

then the probability that household i is observed to have a child in period t is given by

$$\Pr(n_t^i = 1 | M_{t-1}^i) = 1 - \Phi\left(\frac{\xi_t^{i*}}{\sigma}\right), \tag{9}$$

where Φ is the cumulative standard normal. Similarly,

$$\Pr(n_t^i = 0 | M_{t-1}^i) = \Phi\left(\frac{\xi_t^{i*}}{\sigma}\right). \tag{10}$$

The probabilities are conditioned on M_{t-1}^i to emphasize the fact that they depend on past fertility choice. The probability of observing any sequence of birth decisions for a household is merely the product of probabilities such as (9) and (10).

Now, consider the problem of calculating the deterministic part of the decision rules (as given in [8]). To do so requires characterizing the dynamic programming solution explicitly. The exercise is carried out backward through period $T - 1$ only, as the remaining recursions follow by induction.

Expected lifetime utility associated with having a child at $T - 1$ conditional on the information set at $T - 1$ is given by.

$$\begin{aligned}
E_{T-1}(LU_{T-1} | n_{T-1} = 1) &= E_{T-1}(U_{T-1} | n_{T-1} = 1, \xi_{T-1} = 0) \\
&\quad + (P_{T-1} + M_{T-2})\xi_{T-1} + \delta P_{T-1} E_{T-1} \max\, [E_T(LU_T | n_T = 1, \\
&\qquad M_{T-1} = M_{T-2} + 1), E_T(LU_T | n_T = 0, M_{T-1} = M_{T-2} + 1)] \\
&\quad + \delta(1 - P_{T-1}) E_{T-1} \max\, [E_T(LU_T | n_T = 1, M_T, \\
&\qquad E_T(LU_T | n_T = 0, M_{T-1} = M_{T-2})]
\end{aligned}$$

Since ξ_T is not known to the household at t
maximum operators are random. Defining
to be the (unique) value of ξ_T that sets J_T
a child at $T - 1$ who survives ($M_{T-1} = M_T$
on all other information as of $T - 1$, and likewise
$= M_{T-2}$) to be the value of ξ_T that sets $J_T = 0$ conditional either on having a child at $T - 1$ who does not survive or on not having a child at $T - 1$ ($M_{T-1} = M_{T-2}$), and conditional on all other information as of $T - 1$, equation (11) can be written as

$$\begin{aligned}
&E_{T-1}(LU_{T-1} | n_{T-1} = 1) \\
&\quad = E_{T-1}(U_{T-1} | n_{T-1} = 1, \xi_{T-1} = 0) + (P_{T-1} + M_{T-2})\xi_{T-1} \\
&\qquad + \delta P_{T-1}\Big[\Big(\{E_{T-1}E_T(LU_T | n_T = 1, M_{T-1} = M_{T-2} + 1, \xi_T = 0) \\
&\qquad\quad + (P_T + M_{T-2} + 1)E_{T-1}[\xi_T | \xi_T \geq \xi_T^{**}(M_{T-1} = M_{T-2} + 1)]\}
\end{aligned}$$

$$
\begin{aligned}
&\quad \times \{\mathrm{pr}[\xi_T \geqslant \xi_T^{**}(M_{T-1} = M_{T-2} + 1)]\}\Big) \\
&\quad + \Big(\{E_{T-1}E_T(LU_T|n_T = 0, M_{T-1} = M_{T-2} + 1, \xi_T = 0) \\
&\quad + (M_{T-2} + 1)E_{T-1}[\xi_T|\xi_T < \xi_T^{**}(M_{T-1} = M_{T-2} + 1)]\} \\
&\quad \times \{\mathrm{pr}[\xi_T < \xi_T^{**}(M_{T-1} = M_{T-2} + 1)]\}\Big)\Big] \qquad (12)\\
&+ \delta(1 - P_{T-1})\Big[\Big(\{E_{T-1}E_T(LU_T|n_T = 1, M_{T-1} = M_{T-2}, \xi_T = 0) \\
&\quad + (P_T + M_{T-2})E_{T-1}[\xi_T|\xi_T \geqslant \xi_T^{**}(M_{T-1} = M_{T-2})]\} \\
&\quad \times \{\mathrm{pr}[\xi_T|\xi_T \geqslant \xi_T^{**}(M_{T-1} = M_{T-2})]\}\Big) \\
&\quad + \Big(\{E_{T-1}E_T(LU_T|n_T = 0, M_{T-1} = M_{T-2}, \xi_T = 0) \\
&\quad + M_{T-2}E_{T-1}[\xi_T|\xi_T < \xi_T^{**}(M_{T-1} = M_{T-2})]\} \\
&\quad \times \{\mathrm{pr}[\xi_T|\xi_T < \xi_T^{**}(M_{T-1} = M_{T-2})]\}\Big)\Big].
\end{aligned}
$$

In analogous fashion, the expected lifetime utility associated with not having a child at $T - 1$ is

$$
\begin{aligned}
&E_{T-1}(LU_{T-1}|n_{T-1} = 0) \\
&\quad = E_{T-1}(U_{T-1}|n_{T-1} = 0, \xi_{T-1} = 0) + M_{T-2}\xi_{T-1} \\
&\quad\quad + \delta\Big[\Big(\{E_{T-1}E_T(LU_T|n_T = 1, M_{T-1} = M_{T-2}, \xi_T = 0) \\
&\quad\quad\quad + (P_T + M_{T-2})E_{T-1}[\xi_T|\xi_T \geqslant \xi_T^{**}(M_{T-1} = M_{T-2})]\} \\
&\quad\quad\quad \times \{\mathrm{pr}[\xi_T \geqslant \xi_T^{**}(M_{T-1} = M_{T-2})]\}\Big) \qquad (13)\\
&\quad\quad\quad + \Big(\{E_{T-1}E_T(LU_T|n_T = 0, M_{T-1} = M_{T-2}, \xi_T = 0) \\
&\quad\quad\quad + M_{T-2}E_{T-1}[\xi_T|\xi_T < \xi_T^{**}(M_{T-1} = M_{T-2})]\} \\
&\quad\quad\quad \times \{\mathrm{pr}[\xi_T < \xi_T^{**}(M_{T-1} = M_{T-2})]\}\Big)\Big].
\end{aligned}
$$

Under the normality assumption, if we recall that ξ_T is serially uncorrelated, conditional expectations are

$$
\begin{aligned}
E(\xi_T|\xi_T \geqslant \xi_T^{**}) &= \frac{\sigma\phi(\xi_T^{**}/\sigma)}{1 - \Phi(\xi_T^{**}/\sigma)}, \\
E(\xi_T|\xi_T < \xi_T^{**}) &= \frac{-\sigma\phi(\xi_T^{**}/\sigma)}{\Phi(\xi_T^{**}/\sigma)}.
\end{aligned}
\qquad (14)
$$

The decision rule given by (8) for $t = T - 1$ can now be obtained from (12) and (13). The complete dynamic programming solutions can be obtained numerically by continuing to work backward in the fashion described above. Differencing (12) and (13) yields

$$
\begin{aligned}
J_{T-1} = {} & E_{T-1}(U_{T-1} \mid n_{T-1} = 1, \xi_{T-1} = 0) \\
& - E_{T-1}(U_{T-1} \mid n_{T-1} = 0, \xi_{T-1} = 0) + P_{T-1}\xi_{T-1} \\
& + \delta P_{T-1}\Bigg(\{E_{T-1}E_T(LU_T \mid n_T = 1, M_{T-1} = M_{T-2} + 1, \xi_T = 0)\} \\
& \times \left\{1 - \Phi\left[\frac{\xi_T^{**}(M_{T-1} = M_{T-2} + 1)}{\sigma}\right]\right\} \\
& + \{E_{T-1}E_T(LU_T \mid n_T = 0, M_{T-1} = M_{T-2} + 1, \xi_T = 0)\} \\
& \times \left\{\Phi\left[\frac{\xi_T^{**}(M_{T-1} = M_{T-2} + 1)}{\sigma}\right]\right\} \\
& - \{E_{T-1}E_T(LU_T \mid n_T = 1, M_{T-1} = M_{T-2}, \xi_T = 0)\} \\
& \times \left\{1 - \Phi\left[\frac{\xi_T^{**}(M_{T-1} = M_{T-2})}{\sigma}\right]\right\} \\
& - \{E_{T-1}E_T(LU_T \mid n_T = 0, M_{T-1} = M_{T-2}, \xi_T = 0)\} \\
& \times \left\{\Phi\left[\frac{\xi_T^{**}(M_{T-1} = M_{T-2})}{\sigma}\right]\right\} \\
& + P_T\sigma\left\{\phi\left[\frac{\xi_T^{**}(M_{T-1} = M_{T-2} + 1)}{\sigma}\right]\right. \\
& \left. - \phi\left[\frac{\xi_T^{**}(M_{T-1} = M_{T-2})}{\sigma}\right]\right\}\Bigg).
\end{aligned}
\tag{15}
$$

Notice that all terms within the conditional [illegible] utility that are independent of the fertility choice, [illegible] $n_{T-1} = 1$ or 0, will not appear in J_{T-1}. These terms get differenced out, including the square of predicted income and the variance in income; only the conditional mean of income enters the decision rule.[21]

Putting the pieces together, a ξ_t^* is calculated for each t and for a given parameter vector (α_1, α_2, β_1, β_2, γ, c, b, δ) according to the dynamic programming solution outlined above. The likelihood taken over a sample of I women each of whom is observed for T_i periods is

---

[21] Observe that in the previous footnote $(H_t\pi_2)^2$ and σ^2 appear in all expectations regardless of the value of n_t or d_t, a carry-over of the certainty equivalence result in the static model

$$\prod_{i=1}^{I}\prod_{t=1}^{T_i} \mathrm{pr}(n_t^i = \bar{n}_t^i | M_{t-1}^i), \tag{16}$$

where $\bar{n}_t^i$ is one if woman i has a birth at t and zero otherwise. Choosing alternative parameter values leads to alternative critical values (ξ_t^*) and therefore to different sample likelihood values. A derivative-free maximization routine is required in the absence of analytical derivatives. The computational burden lies in the fact that for each set of parameter values the dynamic programming algorithm must be used to calculate the new set of ξ_t^*'s for each individual.

As in standard probit analysis, σ is not identified even though it enters the decision rules explicitly; that is, ξ_t^* is a function of σ. Utility function parameters are thus only identified relative to each other, but cost parameters are identified in an absolute sense.[22] Heuristically, there should be more than sufficient information to estimate the parameters since all future-period expected incomes and survival probabilities enter each period's decision. Further, since ξ_t^* is serially uncorrelated, any set of periods may be used to calculate consistent maximum-likelihood estimates as the current choice is statistically independent of prior choices on which it is conditioned.

Although a generalized error structure for ξ_t is conceptually feasible, it is computationally impractical. A simple permanent-transitory scheme can, however, be estimated and fits neatly into the previous formulation. As is conventional, let $\xi_t^i = \nu^i + \eta_t^i$, where ν^i is a household-specific taste parameter known to the household in the initial period and η_t^i is a purely random taste component. In this case, contrary to the purely transitory taste scheme, since the decision rule at any period t is conditioned on previous fertility decisions as summarized in M_{t-1}, which is itself related to the unobserved permanent taste component, ν^i, estimates will not in general be consistent. However, if the decision is conditioned on the initial nonstochastic state $M_0 = 0$, that is, if the data contain complete life-cycle information, consistent estimates can be obtained.[23]

---

[22] This can be seen from (15). Increasing σ proportionately by λ and likewise the utility function parameters (α_1, α_2, β_1, β_2, γ) by λ leaves ξ_t^{**}/σ unchanged but increases all of the conditional expectations in (15) by λ. Thus, ξ_{t-1}^* increases proportionately by λ while ξ_{t-1}^*/σ is unchanged, which implies that the likelihood is unchanged.

[23] See Heckman (1979) for a discussion of the problem of initial conditions in discrete models and for suggested solutions. Notice too that there is true state dependence, although there is no single structural state dependence parameter. If identification of the structural parameters of the model is not problematic, then there is also no difficulty in disentangling true from spurious state dependence.

Suppose then that there are N types of households ($N < I$) with permanent taste components equal to $v^1, \ldots, v^N$ and comprising $q^1, \ldots, q^n$ proportions of the population, respectively. For each of the N possible values of v^i one can compute ξ_t^* for each t exactly as was done in the single type case. One can therefore form the likelihood of observing any birth sequence given household type. Weighting a given sequence's likelihood for each type by the appropriate probabilities q^i, summing over all types, and producting over all individuals yields the sample likelihood. The distribution of the unobservable is then estimated along with other parameters, including possibly N.[24] The computational cost of introducing unobserved heterogeneity is approximately proportional to the number of household types for known N.

Data

The 1976 Malaysian Family Life Survey contains 1,262 households consisting of at least one ever-married woman under 50 years of age as of the survey date. The essential feature of the survey is that it contains a retrospective life history of each woman to age 15 or marriage, whichever is earlier, with complete information about births, infant deaths, and household income. A subsample was drawn consisting of 188 Malay women, all of whom were over 30, currently married, and married only once.[25] Household income is restricted to that of the husband.[26]

The length of the period was chosen to b
mise between the necessity to have only a s
and the computational benefit from havin
cycle.[27] The initial period was set at ag
whichever came first. The final decision p
after 20 periods or approximately at age 1

In treating husband's income (earnings), the
was broken up into 18-month periods and matc
life cycle. Starting with the initial period (age

---

[24] See Heckman and Singer (1982) for a discussion of nonp
estimating unobserved heterogeneity distributions in the context of continuous time duration models.

[25] Approximately 50 percent of the 1,262 women are Malay the other 50 percent consisting of ethnic Chinese and Indians

[26] Essentially, it is assumed that the woman does no market work or work at home for pay Including the woman's income would have led to negative income effects since women who work for pay have fewer children. A weakness of the model is that it does not consider the female labor supply choice

[27] Of the total of 3,086 periods, in only 30 periods (27 women) were there two births in 18 months. In those cases, a birth was moved to the next period in which there was no birth.

dynamic programming solution outlined above. The likelihood taken over a sample of I women each of whom is observed for T_i periods is

$$\prod_{i=1}^{I}\prod_{t=1}^{T_i} \mathrm{pr}(n_t^i = \bar{n}_t^i | M_{t-1}^i), \tag{16}$$

where $\bar{n}_t^i$ is one if woman i has a birth at t and zero otherwise. Choosing alternative parameter values leads to alternative critical values (ξ_t^*) and therefore to different sample likelihood values. A derivative-free maximization routine is required in the absence of analytical derivatives. The computational burden lies in the fact that for each set of parameter values the dynamic programming algorithm must be used to calculate the new set of ξ_t^*'s for each individual.

As in standard probit analysis, σ is not identified even though it enters the decision rules explicitly; that is, ξ_t^* is a function of σ. Utility function parameters are thus only identified relative to each other, but cost parameters are identified in an absolute sense.[22] Heuristically, there should be more than sufficient information to estimate the parameters since all future-period expected incomes and survival probabilities enter each period's decision. Further, since ξ_t^* is serially uncorrelated, any set of periods may be used to calculate consistent maximum-likelihood estimates as the current choice is statistically independent of prior choices on which it is conditioned.

Although a generalized error structure for ξ_t is conceptually feasible, it is computationally impractical. A simple permanent-transitory scheme can, however, be estimated and fits neatly into the previous formulation. As is conventional, let $\xi_t^i = \nu^i + \eta_t^i$, where ν^i is a household-specific taste parameter known to the household in the initial period and η_t^i is a purely random taste component. In this case, contrary to the purely transitory taste scheme, since the decision rule at any period t is conditioned on previous fertility decisions as summarized in M_{t-1}, which is itself related to the unobserved permanent taste component, ν^i, estimates will not in general be consistent. However, if the decision is conditioned on the initial nonstochastic state $M_0 = 0$, that is, if the data contain complete life-cycle information, consistent estimates can be obtained.[23]

---

[22] This can be seen from (15). Increasing σ proportionately by λ and likewise the utility function parameters (α_1, α_2, β_1, β_2, γ) by λ leaves ξ_t^{i*}/σ unchanged but increases all of the conditional expectations in (15) by λ. Thus, ξ_{t-1}^* increases proportionately by λ while ξ_{t-1}^*/σ is unchanged, which implies that the likelihood is unchanged.

[23] See Heckman (1979) for a discussion of the problem of initial conditions in discrete models and for suggested solutions. Notice too that there is true state dependence, although there is no single structural state dependence parameter. If identification of the structural parameters of the model is not problematic, then there is also no difficulty in disentangling true from spurious state dependence.

Suppose then that there are N types of households ($N < I$) with permanent taste components equal to $v^1, \ldots, v^N$ and comprising $q^1, \ldots, q^n$ proportions of the population, respectively. For each of the N possible values of v^i one can compute ξ_t^* for each t exactly as was done in the single type case. One can therefore form the likelihood of observing any birth sequence given household type. Weighting a given sequence's likelihood for each type by the appropriate probabilities q^i, summing over all types, and producting over all individuals yields the sample likelihood. The distribution of the unobservable is then estimated along with other parameters, including possibly N.[21] The computational cost of introducing unobserved heterogeneity is approximately proportional to the number of household types for known N.

Data

The 1976 Malaysian Family Life Survey contains 1,262 households consisting of at least one ever-married woman under 50 years of age as of the survey date. The essential feature of the survey is that it contains a retrospective life history of each woman to age 15 or marriage, whichever is earlier, with complete information about births, infant deaths, and household income. A subsample was drawn consisting of 188 Malay women, all of whom were over 30, currently married, and married only once.[25] Household income is restricted to that of the husband.[26]

The length of the period was chosen to
mise between the necessity to have only a
and the computational benefit from havin
cycle.[27] The initial period was set at age
whichever came first. The final decision
after 20 periods or approximately at age

In treating husband's income (earnings), the
was broken up into 18-month periods and matched to the woman's life cycle. Starting with the initial period (age) of the woman, hus-

---

[21] See Heckman and Singer (1982) for a discussion of nonparametric methods for estimating unobserved heterogeneity distributions in the context of continuous time duration models

[25] Approximately 50 percent of the 1,262 women are Malay, the other 50 percent consisting of ethnic Chinese and Indians

[26] Essentially, it is assumed that the woman does no market work or work at home for pay. Including the woman's income would have led to negative income effects since women who work for pay have fewer children. A weakness of the model is that it does not consider the female labor supply choice.

[27] Of the total of 3,086 periods, in only 30 periods (27 women) were there two births in 18 months In those cases, a birth was moved to the next period in which there was no birth.

band's (average) income (possibly prospective if they were not then married) was computed for each 18-month period. In the data themselves husband's income is reported in an event-history form; namely, labor supply and wage rates are reported retrospectively at each moment (date) that either of them changes beginning at husband's age 15. A semilog (price-deflated) earnings function was estimated for each husband individually with a constant, period, and period squared as regressors, where period is simply age of husband suitably transformed to the woman's life cycle.[28] A predicted earnings profile was generated and forms the basis for the (forecasted) income observations used in the empirical analysis.[29] The woman is assumed to live for 10 periods (15 years) after her childbearing period, and husband's income for the period is assumed to be forecasted as constant and equal to predicted income in the twenty-first period. In the context of the model of the previous section, it is assumed that each woman knows the parameters of her husband's earnings functions beginning at the first decision period.

The survival probability was obtained from data on state-specific survival rates. Predicted survival probabilities are based on state-specific log odds regressions, which include a constant, time, and time squared. These predictions are assumed to be those used by households in making their fertility decisions. Less aggregated data would have been more appropriate but are not available, and the sample size is too small to do anything reasonable with the individual data.[30]

The data therefore consist of the actual periods of births and infant deaths for each woman, a predicted husband's income profile based on a semilog quadratic in time formulation, and a predicted survival probability profile based on a quadratic in time-logistic formulation. In each period, the probability of a birth depends in a complicated way on the number of surviving children to that time, the predicted future income path of the husband, and the predicted future survival probability path.

Estimation Results

The model was estimated for two samples, one using the birth outcomes of the first 10 periods of each woman's life cycle (1,880 periods) and the other using birth outcomes over each woman's entire life cycle (3,086 periods). The 10-period sample was used to test for unobserved heterogeneity since the composition of the sample changes as more periods are added. Because of computational cost limitations, only a model with two types of individuals was estimated. Since a likelihood ratio test clearly rejected the existence of unobserved heterogeneity, results from the 10-period sample are not reported.[31] It is assumed, therefore, that unobserved heterogeneity will not, if ignored, seriously contaminate estimates using the entire life cycle of each woman.[32] The estimates presented below therefore are based on the assumption that the random taste component of the model follows a serially independent stochastic process.

---

[28] Earnings are deflated by the countrywide price index that prevailed at the beginning of each 18-month period. It is obtained from *National Accounts of West Malaysia, 1947–76*

[29] Notice that I abstract from any serial correlation in the earnings function. New information acquired during the life cycle requires updating the future component of the dynamic programming decision rule each period, which greatly increases the computational burden.

[30] It was necessary to extrapolate outside the period covered by the state-level data Also, the state-level data did not precisely conform to the data that were required. Details are provided in Wolpin (1982)

The model is formulated slightly more generally than in previous sections. In particular, the costs of birth in the first and second periods are permitted to differ from one another and from subsequent periods, and the cost profile is permitted to have a quadratic shape. It is unlikely a priori that income and survival probability profiles can by themselves trace the life-cycle fertility profile, as shown in table 1. Moreover, presumed biological constraints early and late in the life cycle associated with a reduced propensity to conceive can be formally expressed as a reduced contraceptive cost or an increased net cost of a birth. In addition, the woman's schooling is arbitrarily introduced as affecting the incremental utility of a surviving child.[33]

The structure of the model is, therefore, as follows.

$$LU = \sum_{t=1}^{T} \delta^{t-1}(\alpha_{1t}M_t - \alpha_2 M_t^2 + \beta_1 X_1 - \beta$$

$$Y_t = X_t + b(n_t - d_t) + (c_1 m_1 + c_2 m_2$$

where, in addition to previous definitions

equal to unity if the period is the first a

---

[31] Twice the difference in the log likelihood values of the two models was only 2, which falls far short of the critical $\chi^2(.05) = 5.99$. Of course, having only two types of individuals might not capture the true degree of heterogeneity in the sample.

[32] There is other evidence to support this view. First, the observed variance in the number of children born taken over all women and all periods is 1,391, the variance that would be obtained if the birth process was simply Bernoulli is 733, and the variance that would obtain if women had either a child each period or no children at all is 11,984, which suggests that extreme heterogeneity is not present. Second, there seems to be little discernible relationship between the number of surviving children and the observed frequency of children born in the subsequent period. The model predicts a negative relationship, and unobserved heterogeneity a positive relationship. However, both of these pieces of evidence abstract from observable heterogeneity determinants.

[33] Schooling could be modeled as affecting any or all cost or utility parameters. Since it is unlikely that these alternatives could be distinguished empirically, the simplest representation was adopted.

TABLE 1

ACTUAL AND PREDICTED BIRTH PROBABILITIES

	PERIOD									
	1	2	3	4	5	6	7	8	9	10
Actual	138	335	.489	537	.511	.505	511	516	516	.394
Predicted	.134	330	543	538	.530	518	502	.483	.460	.434
	11	12	13	14	15	16	17	18	19	20
Actual	410	.425	321	360	310	262	200	.140	140	.105
Predicted	.406	382	349	316	.282	249	213	.178	141	.112

dummy variable equal to unity if the period is the second and zero otherwise, and S is the woman's years of schooling. Thus $c_1 + c_{30} + c_{31} + c_{32}$ measures the birth cost in period 1, $c_2 + c_{30} + 2c_{31} + 4c_{32}$ the birth cost in period 2, and $c_{30} + c_{31}t + c_{32}t^2$ the birth costs in all periods after the second.

Table 2 presents the estimated parameters of the model.[34] Although individual parameters cannot be translated into the experiments of interest, they do provide a simple check on the plausibility of the estimates, and thus of the method. All of the signs in table 2 conform to strong priors; incremental or marginal utilities are positive and diminishing in surviving children and in the composite good, the cost of bearing a child is positive in all life-cycle periods as is the maintenance cost, and the discount factor translates into a rate-of-time preference of approximately .093. Magnitudes of the cost parameters are also not unreasonable, though at the high end. The cost of a live birth in period 1 (c_1) is M$22,280 (Malaysian dollars) and in period 2 (c_2) M$8,121. The cost of a birth is lowest in period 3, M$3,314, and rises at an increasing rate reaching M$25,407 by period 20. These costs are high given that the average level of husband's income was about M$12,000 and that the Malaysian dollar was worth about 33 percent of the U.S. dollar in 1960.[35] The child-maintenance cost, on the other hand, is relatively small, only M$1,470. Notice also

---

[34] Two maximization routines were used to obtain the final estimates. First, a search routine NMSIMP was used to scan broadly over the parameter space and then DFP was used to narrow in on the maximum. Both routines are part of the GQOPT package. Asymptotic standard errors are not reported since the hessian (at the maximum) was not negative definite. The model was also estimated with $c_{31} = c_{32} = 0$. The restriction was rejected by the appropriate likelihood ratio test. The χ^2 value for the likelihood ratio test was 204, which exceeds the critical $\chi^2_2(.05) = 5.99$.

[35] The quadratic utility specification does not rule out the possibility that, because of the magnitude of some of these cost parameters, consumption will be negative at some period for some individual.

TABLE 2

MAXIMUM LIKELIHOOD ESTIMATES

Utility Function Parameters		Cost Function Parameters	
α_1 =	3.43×10^{-2}	b =	1.47×10^{3}
α_2 =	2.94×10^{-7}	c_1 =	1.77×10^{4}
β_1 =	6.16×10^{-5}	c_2 =	8.09×10^{3}
β_2 =	1.07×10^{-16}	c_{30} =	1.95×10^{3}
γ_1 =	2.42×10^{-7}	c_{31} =	-2.86×10^{2}
γ_2 =	-5.42×10^{-3}	c_{32} =	6.42×10^{1}
δ =	9.22×10^{-3}		

NOTE.—Log likelihood = −1,920.3.

that the incremental utility of an additional surviving child rises with goods consumption ($\gamma_1 > 0$) and falls with mother's schooling ($\gamma_2 > 0$). However, quantitatively the nonlinear terms in the utility function other than schooling are quite small, and this feature is crucial to the substantive implications discussed below.

There are several ways to assess the fit of the estimated model. The log likelihood value as shown in table 2 is −1,920.3. The model degenerates to a simple Bernoulli process for births if all of the parameters except α_1 are set to zero; α_1 must be estimated to retrieve the sample fraction of births. The log likelihood value for the pure chance model is −2,059.9. Twice the difference in the likelihood is 272.4, which is sufficient to reject the pure chance model at almost any significance level ($\chi^2[.01] = 26.2$ with 1

Table 1 compares the actual and
profiles. Although they are quite s
specification of childbearing costs. If th
are collapsed to a single parameter, th
profile is essentially flat rather than
more revealing comparison is shown in
predicted birth probabilities conditional on
children are shown. Tests for the existence of a relationship by period in the actual data are shown by the first row of χ^2 statistics. The second row of statistics tests for the equality between predicted and actual. Notice first that the manner in which the actual birth probabilities vary with the number of surviving children differs across periods, where there is a statistically significant relationship (periods 8, 10, and 12), at least in one the relationship appears quadratic, the probability first rising, then falling. In only one period is there a discernible difference between the actual and the predicted probabilities, and there it seems to be the level that is misrepresented and not the

---

[10] Such a specification still does better than the pure chance model

TABLE 3

PROBABILITY OF A BIRTH CONDITIONAL ON THE NUMBER OF SURVIVING CHILDREN, ACTUAL AND PREDICTED

NUMBER OF SURVIVING CHILDREN	PERIOD																							
	3		4		5		6		7		8		9		10		11		12		13		14	
	A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P	A	P
0 or 1	486	464	533	539	477	523	507	507	538	481	389	472	435	391			438	375						
1 or 2	549	454	604	622	589	540	509	528	558	500	500	471	408	408	379	400	367	400	292	375			133	333
2 or 3			556	555	544	558	514	522	495	514	529	480	470	458	375	422	275	392	209	372	256	307	241	310
3 or 4					555	550	532	511	429	519	505	495	570	470	451	439	377	406	320	380	225	325	289	.316
4 or 5									500	500	543	500	594	464	442	442	430	407	574	382	220	341	369	316
5 or 6													519	444	415	434	435	406	574	382	323	354	400	320
6 or 7															550	444	552	414	459	378	379	362	340	321
7 or 8															440	444	533	400	350	232	483	379	333	333
8 or 9																					500	417	474	369
χ^2 (actual)	813		1.30		2.95		1.52		4.77		10.7*		6.63		14.5*		11.1		28.9*		16.3		18.8	
χ^2 (difference)	2.20		1.34		2.11		659		5.39		10.38		9.60		13.02		8.08		29.22*		12.21		15.57	

NOTE.—Cells with fewer than five observations are omitted from the total.

* Indicates statistically significant at the 5 percent level.

pattern. Thus, if there exists a relationship between the propensity to bear a child in any period and prior net fertility, it is complex and the estimated model seems to capture its essential features.

The estimated parameters may be used to explore the sensitivity of fertility over the life cycle to changes in the income profile, the survival probability profile, child deaths, and mother's schooling.

Replacement Effects

Table 4 shows the responsiveness of fertility to the occurrence of infant deaths. Each entry gives the expected number of children ever born (sums of unconditional probabilities) given that there are M_{t-1} surviving children as of period $t = 2, \ldots, 10$. Average (constant over time) income and survival probabilities and average schooling are assumed in the calculation. Replacement responses can easily be calculated by differencing any two columns for a given row. The last column averages the replacement effects within each row. As the figures reveal, replacement responses are small for the sample, an infant death induces an increase in the number of children ever born by at most 0.015.[37]

Income Effects

Income effects are also quite small. Elasticities tend to rise with the level of income with arc elasticities ranging from [illegible] between $5,000 and $10,000 to .06 for [illegible] $50,000 (table available on request). Also [illegible] to increase births more in earlier life-cycle [illegible] come profile tends to delay births slightly [illegible]

Survival Probability Effects

Survival probability effects are not small. A reduction by [illegible] percentage points reduces the number of children ever born by about one-quarter (table available on request). As the survival probability falls, there is also a tendency to have children earlier. A rising survival probability profile tends to delay childbearing. Thus, women are more likely to bear children in periods when they are more likely to

---

[37] Olsen (1983) finds replacement effects for these survey data (but not the same sample) of about 30 percent. His methodology is "statistical" in nature.

TABLE 4

Expected Additional Number of Births by Number of Surviving Children and Period

Period	Number of Surviving Children 0	1	2	3	4	5	6	7	8	9	Average Replacement
2	6.579	6.566									.013
3	6.229	6.214	6.200								.015
4	5.675	5.661	5.650	5.638							.012
5	5.130	5.000	4.987	4.974	4.964						.012
6	4.595	4.583	4.571	4.561	4.549	4.537					.012
7	3.868	3.856	3.844	3.832	3.819	3.810	3.798				.011
8	3.388	3.377	3.367	3.355	3.346	3.334	3.323	3.312			.011
9	2.937	2.928	2.917	2.907	2.896	2.884	2.876	2.865	2.854		.010
10	2.648	2.638	2.628	2.618	2.609	2.600	2.590	2.579	2.570	2.560	.010

survive and will bear more children overall in high survival environments.

Mother's Schooling Effect

Mother's schooling exerts a strong effect on fertility. Each additional year of schooling reduces the expected number of children by 0.35 evaluated at the means of the other variables.

Discussion

There is a straightforward interpretation of the estimation results. Essentially, the estimates imply that this sample of Malay women will follow a maximal fertility strategy in the following sense. If the cost of childbearing and the infant survival probability were constant over the life cycle, then each woman would either have a child in each period or have no children at all (subject to the randomly drawn taste parameter each period). The fact that childbearing costs are estimated to be rising over the life cycle limits the total number of children but not the pattern. This result is directly tied to the very negligible magnitudes of the second-order utility function parameters (α_2, β_2, γ_2). There is no replacement behavior because the marginal child, the last child for whom a "positive" net benefit is derived, does not change with child deaths. This would not be the case if the "marginal" utility of surviving children were falling sub

tional surviving children. Moreover, syste

random tastes) cannot arise given the util

mates. Rising income profiles and survival

otherwise generate spacing, but evidently

be consistent with observed fertility profi

the level of fertility is related to the wom

infant mortality environment.

Two issues arise with respect to an assessment of this approach. First, would an astructural approach yield qualitatively similar results, and second, if it did, would the value of the present approach then be merely aesthetic, in the sense that estimation is tied to an optimization framework?

With respect to the second, in order to appreciate the economy of this methodology, one need only compare the number of parameters that would have to be estimated to achieve the same wealth of empirical implications using an astructural methodology. A rough calculation indicates that a minimum of 400 parameters would have to be estimated to obtain all of the responses that the 13-parameter structural model can characterize.[38] If the implicit restrictions of the structural model are correct, a great deal of statistical precision is gained. If incorrect, that precision is illusory. Unfortunately, the lack of analytical solutions for the decision rules seems to preclude a test of the restrictions.

To compare these results to a less structured approach, probit functions were estimated separately for each period of the woman's life cycle. To mirror the structural estimates qualitatively, explanatory variables consisted of the number of surviving children, the current period and 5- and 10-year-ahead expected income levels, the current period and 5- and 10-year-ahead survival probabilities, and mother's schooling. To summarize those results: income effects were negligible and mother's schooling "significantly" reduced fertility, as was found in the structural estimation results. Prior net fertility either had no "significant" effect or was positively related to the current decision. The latter result is not consistent with estimates from the structural model and suggests the existence of unobserved heterogeneity, a simple version of which was rejected in the structural estimation. Finally, the probit estimates imply that a general increase in the survival probability reduces, rather than increases, fertility, although as in the structural estimates the probability of a birth generally decreases as future survival probabilities increase. The probit estimates are thus somewhat different from the structural model estimates, and if the latter are a correct characterization of behavior, the former would also be quite misleading.

Conclusions

This paper is the first attempt specifically to model sequential life-cycle individual fertility behavior in an uncertain environment in a way that is conducive to estimation. More generally, it has demonstrated the empirical feasibility of estimating a dynamic programming model when the problem admits only of numerical solution. However, the theoretical formulation contained many important simplifying assumptions, and the sensitivity of the results to those assumptions is an important issue yet to be addressed. Significant computational issues must be resolved if extensions are to be pursued. Extensions of the basic framework to incorporate endogenous female labor supply, savings, and imperfect contraception would make up a rather full agenda for future research.

---

[38] In the first period, there are 20 future income variables and 19 future survival probabilities In period 2, those numbers are 19 and 18, respectively, plus M_{t-1}. Following this logic for 20 periods yields the figure in the text

The particular problem addressed in this paper, that of the timing and spacing of children given significant infant mortality, has been of concern to many social science researchers. Specifically, the "replacement effect" estimated from this model, within a unified framework of the life-cycle fertility decision, is much smaller than that obtained using other methods. At this stage, in my view, these estimates should be taken no more or less seriously than those based on a looser theoretical tradition.

References

Bellman, Richard E. *Dynamic Programming* Princeton, N J. Princeton Univ. Press, 1957.

Ben-Porath, Yoram. "Fertility Response to Child Mortality. Micro Data from Israel." *J.P.E.* 84, no. 4, pt. 2 (August 1976): S163–S178.

Flinn, Christopher J., and Heckman, James J "New Methods for Analyzing Structural Models of Labor Force Dynamics" *J Econometrics* 18 (January 1982): 115–68.

Hansen, Lars Peter, and Sargent, Thomas J "Linear Rational Expectations Models for Dynamically Interrelated Variables." In *Rational Expectations and Econometric Practice*, edited by Robert E. Lucas, Jr., and Thomas J. Sargent. Minneapolis: Univ Minnesota Press, 1981

Hartley, Michael. "Neoclassical Econometrics, Part One: General Considerations." Washington: World Bank, Development and Research Department Economics and Research Staff, 1981.

Heckman, James J. "Statistical Models for Discrete
graphed. Chicago: Univ. Chicago, January 19

Heckman, James J., and MaCurdy, Thomas
male Labour Supply" *Rev Econ Studies* 47

Heckman, James J., and Singer, Burton
Econometric Models for Duration Data"
Chicago. National Opinion Res. Center,

Heckman, James J., and Willis, Robert J "Estim
Reproduction: An Econometric Approach
Consumption, edited by Nestor E Terleckyj.
Press (for N B E R.), 1976.

Hotz, V. Joseph. "A Life Cycle Model of Fertility and Married Women's Labor Supply." Mimeographed. Pittsburgh: Carnegie-Mellon Univ., October 1980.

Miller, Robert. "An Estimate of a Job Matching Model" Mimeographed Pittsburgh: Carnegie-Mellon Univ., October 1982.

Moffitt, Robert. "Life Cycle Profiles of Fertility A State Dependent Multinomial Probit Model." Mimeographed. New Brunswick, N.J. Rutgers Univ., August 1980.

Olsen, Randall J. "Estimating the Effects of Child Mortality on the Number of Births." *Demography* 17 (November 1980): 429–43

——— "Mortality Rates, Mortality Events, and the Number of Births" *A.E.R.* 73 (May 1983): 29–32.

Pakes, Ariel. "Patents as Options: Some Estimates of the Value of Holding European Patent Stocks." Mimeographed. Jerusalem: Hebrew Univ., December 1983.

Sargent, Thomas J. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E.* 86 (December 1978): 1009–44.

Schultz, T. Paul. "Interrelationships between Mortality and Fertility." In *Population and Development: The Search for Selective Interventions*, edited by Ronald G. Ridker. Baltimore: Johns Hopkins Univ. Press (for Resources for the Future), 1976.

Vijverberg, Wim. "Labor Supply and Fertility Decisions: A Dynamic Model of the Economic Behavior of Married Women." Ph.D. dissertation, Univ. Pittsburgh, 1981.

Wolpin, Kenneth I. "An Estimable Dynamic Model of Fertility and Child Mortality." Mimeographed. New Haven, Conn.: Yale Univ., October 1982.

Cross-Regime Evidence of Macroeconomic Rationality

Roger C. Kormendi and Philip G. Meguire

University of Chicago

Rational expectations macromodels predict that the short-run effects of monetary shocks on real output (χ) should be negatively related across policy regimes to the variability of such shocks. This paper presents cross-regime tests of this and related propositions based on a sample of 47 countries. The within-regime estimates reveal a consistent pattern of positive short-run real output effects, with neutrality of money holding in the long run. The cross-regime tests show that χ is negatively related to the variance of monetary shocks, positively related to the variance of real output shocks, negatively related to the variance of velocity shocks, and unrelated to either the mean or variance of anticipated money growth.

I. Introduction

In his seminal contributions to rational
roeconomic behavior, Robert Lucas (1972
proposition that the effects of aggregate
output should be inversely related to the
bances (see also Lucas 1975; Barro 1976). In
proposition, Lucas (1973) regressed detrended real income on the changes in nominal income (his measure of aggregate nominal disturbances) for each of 18 countries, made an informal cross-country comparison of the resulting coefficients with the variance of each country's inflation rate, and found some support for a negative rela-

We wish to acknowledge a special debt to Robert Barro for his help and guidance on this paper. We also wish to thank Lars Hansen, James Heckman, Bennett McCallum, Michael Mussa, Richard Parks, Roger Waud, and members of the Money Workshop at the University of Chicago for their comments on earlier drafts.

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0014$01.50

tion.[1] This evidence is generally interpreted as rejecting policy-invariant macromodels, that is, models in which the reduced-form trade-off between nominal shocks and real output is *invariant* with respect to the parameters of the process governing those nominal shocks. Unfortunately, because nominal income is *identically* real income times the "price level," the negative cross-country relation uncovered by Lucas can be shown to be predicted by a class of policy-invariant Phillips curve models as well (see Sec. II).

The purpose of this paper is to test Lucas's proposition concerning the relation between the variance of nominal shocks and the effects of such shocks on real output using procedures that avoid this problem with Lucas's original test. In other words, we wish to provide a test capable of rejecting the class of policy-invariant macromodels in favor of the class of rational expectations macromodels. To this end, we combine Barro's (1977, 1978) idea of estimating the effects of unanticipated money supply changes on real output with Lucas's idea of drawing inferences from a cross-section of policy regimes.[2] Section II contains a critical review of Lucas's (1973) test and a preliminary discussion of how it can be reformulated. Section III contains the details of our methodology. In Section IV, we report the estimated decomposition of money supply changes for each country in our sample into anticipated and unanticipated components. In Section V, we estimate the effects of unanticipated money supply changes on real output. In Section VI, we implement cross-regime tests of the Lucas proposition and other related propositions. The evidence rejects policy-invariant macromodels in favor of rational expectations macromodels. Section VII concludes the paper.

II. Testing the Basic Rational Expectations Proposition across Regimes

Lucas (1973) fit the following equation to annual data for each of 18 countries:

$$y_{c,t} = a_0 + a_1 \Delta x_t + a_2 y_{c,t-1} + u_t, \tag{1}$$

where $y_{c,t}$, the "cyclical" component of (log) real output, is the residual from an estimate of $y_t = c_0 + c_1 t + y_{c,t}$, and Δx_t is the change in (log) nominal income (which he takes as his measure of nominal shocks). Lucas then asked whether his estimates of $\hat{a}_1$ were negatively related to his estimates of the variance of Δx, $\hat{\sigma}_x^2$. Comparing $\hat{a}_1$ and $\hat{\sigma}_x^2$ across the 18 countries showed that the two countries with the greatest $\hat{\sigma}_x^2$ (Argentina and Paraguay) had much smaller $\hat{a}_1$ than the 16 remaining countries. Unfortunately, an important class of alternative models predicts the same negative relationship between $\hat{a}_1$ and $\hat{\sigma}_x^2$ even when expectations are not rational.

---

[1] Alberro (1981) extended Lucas's test to a sample of 50 countries and obtained similar results. See also Froyen and Waud (1980) for related and more detailed tests over 10 industrialized countries.

[2] Barro (1977, 1978), Mishkin (1982), and others test another implication of rational expectations macromodels—that anticipated money supply changes have no real output effect whereas unanticipated monetary shocks do have such effects. Such tests require within-regime identifying restrictions. As Sargent (1976*b*, p. 637) has noted, even if one cannot achieve the within-regime identification required for such tests, *cross-regime* tests of the relation between monetary variance and the real output effects of money shocks are still feasible (see also McCallum 1979 and Weber 1981).

Consider the following policy-invariant ad hoc Phillips curve–type equation:

$$y_{c,t} = b_0 + b_1\Delta p_t + b_2 y_{c,t-1} + e_t, \tag{2}$$

where Δp_t is the change in the (log) price level, e_t is a white-noise error term, and b_1 is a constant that is not functionally related to σ_p^2, that is, $\partial b_1/\partial\sigma_p^2 = 0$.[3] Assume also that e_t, Δp_t, and $y_{c,t-1}$ are mutually orthogonal. Using the transformation $\Delta y_t = c_1 + \Delta y_{c,t}$, equations (1) and (2) can then be rewritten as

$$\Delta y_t = (a_0 + c_1) + a_1\Delta x_t + (a_2 - 1)y_{c,t-1} + u_t \tag{1'}$$

and

$$\Delta y_t = (b_0 + c_1) + b_1\Delta p_t + (b_2 - 1)y_{c,t-1} + e_t. \tag{2'}$$

Now consider estimating (1′) when (2′) is in fact true. Substituting (2′) into the identity $\Delta x_t \equiv \Delta y_t + \Delta p_t$ yields

$$\Delta x_t = (b_0 + c_1) + (1 + b_1)\Delta p_t + (b_2 - 1)y_{c,t-1} + e_t. \tag{3}$$

Using (2′) and (3), we can solve for the (po [illegible]
a_1 in (1′) as

$$\hat{a}_1 = \frac{b_1(b_1 + 1)\sigma_p^2 + ([illegible]}{(b_1 + 1)^2\sigma_p^2 + (b [illegible]}$$

where σ_e^2 is the variance of e_t in (2′) an [illegible]
$\partial\hat{a}_1/\partial\sigma_p^2 < 0$, a negative relation between [illegible]
variation in σ_x^2 across countries is dominated [illegible]
is true in Lucas's sample. Therefore, Lucas's empirical results are implied by (2), even though such an equation assumes a policy-invariant Phillips curve and no rationality of expectations.[4]

One can avoid this problem with Lucas's test by employing a real output equation such as Barro's, in which the crucial independent variables are current and lagged values of unanticipated money supply changes. Doing so breaks the identity linking the dependent and independent variables in the real output equation and hence eliminates the source of the problem. Consider, then, the following simple variant of Barro's real output equation:

---

[3] One could alternatively structure a policy-invariant *expectational* Phillips curve such as $y_{c,t} = b_0 + b_1(\Delta p_t - \Delta p_t^e) + b_2 y_{c,t-1} + e_t$, where Δp_t^e is the expectation of Δp_t based on information available at $t - 1$, and the conclusion in the text would be essentially unaltered.

[4] Although we have assumed for simplicity that Δp_t and $y_{c,t-1}$ are orthogonal, relaxing that assumption does not affect the main result that $\partial\hat{a}_1/\partial\sigma_p^2 < 0$. See Barro (1981) for a similar analysis.

$$DY_t = \beta(L)RM_t + \gamma(L)DY_{t-1} + RY_t. \qquad (5)$$

RM_t is the unanticipated money supply change in period t. DY_t is the change in (log) real output, assumed to depend on a distributed lag of RM_t and perhaps some own lagged values. The OLS projection coefficients $\beta(L)$ and $\gamma(L)$ are defined as $\Sigma_{i=0}^{N_1} \beta_i L^i$ and $\Sigma_{i=1}^{N_2} \gamma_i L^i$, respectively, where L is the lag operator and N_1 and N_2 are assumed for empirical purposes to be finite. RY_t is the OLS white-noise residual. (A constant term has been omitted without loss of generality.) In the context of (5), Lucas's proposition holds that the magnitude of the effects of RM_t on DY_t embodied in the β_i coefficients should be related to the variance of RM_t, VRM. In contrast, the class of policy-invariant macromodels admits of no relation between the β_i and VRM. Thus, investigating whether the β_i are related to VRM in a systematic way across policy regimes provides the basis for tests capable of rejecting the class of policy-invariant macromodels in favor of rational expectations models.

Some complications arise in structuring tests of Lucas's proposition that deserve discussion at this point. The β_i coefficients in an equation such as (5) are properly interpreted as reduced-form (or, as Sargent [1979] has put it, "mongrel") coefficients. Thus, in addition to any direct structural effects of monetary shocks on real output, the β_i may incorporate the effects of omitted (and perhaps unobservable) state variables. In this regard, even if the direct structural effects of RM_t on DY_t are contemporaneous, the reduced-form β_i coefficients in (5) may involve (potentially quite complex) lagged effects, in which case simple predictions regarding the effects of VRM on the β_i are not generally obtainable (see Sargent 1976*b*, 1979; McCallum 1979; Kormendi and Meguire 1982). Despite these complications, what might be termed the "weak form" of Lucas's proposition—that the β_i are related (but not necessarily negatively) to VRM—can still be tested against the nested null hypothesis that the β_i are invariant with respect to VRM. Our interest, however, will be in a stronger form of Lucas's proposition to be discussed shortly.

Although the β_i can in principle reflect a complex structure of lagged effects of RM_t on DY_t, there are some properties that, intuitively, one might expect to characterize the β_i. First, one might expect the "short-run" real output effect of monetary shocks to be nonnegative. Second, one might expect monetary shocks to be neutral in the "long run." Third, one might expect the "peak effect" of a monetary shock on real output to occur within at most two years of the shock itself. We can characterize these three properties in terms of the notation of equation (5) as follows.

Since equation (5) involves the first difference of Y_t, the impact of a unit RM_t shock on $Y_{t+\tau}$, which we denote by χ_τ, is the sum of the first τ β_i coefficients, that is, $\chi_\tau = \Sigma_{i=0}^{\tau} \beta_i$. If monetary shocks have nonnegative effects on $Y_{t+\tau}$ in the "short run," that is, for $\tau \leq N_1$, then

$$\chi_\tau \equiv \sum_{i=0}^{\tau} \beta_i \geqslant 0 \quad \text{for } \tau = 0, 1, \ldots, N_1. \tag{P1}$$

If monetary shocks are long-run neutral, then the effect of RM_t on Y_{t+N_1} would be zero, that is,

$$\chi_{N_1} \equiv \sum_{i=0}^{N_1} \beta_i = 0. \tag{P2}$$

Finally, the peak effect of an RM_t shock on real output can be defined as

$$\chi = \max_{\tau} \{\chi_\tau\}, \quad \tau = 0, 1, \ldots, N_1. \tag{6}$$

The third property is simply

$$\chi_\tau = \chi \text{ at } \tau \leq 2 \tag{P3}$$

when (5) is specified with annual data.[5] These [illegible]
give rise to β_i like those in figure 1. The [illegible]
in figure 1, would then have the "tr[illegible]
level specification of Barro (1978) and [illegible]

We can now characterize the "stron[illegible]
that will be the main focus of the pape[illegible]
"peak effect" of a monetary shock on [illegible]
to the variability of the RM_t shocks. That [illegible]
form rational expectations hypothesis that χ a[illegible]
related against the nonrational null hypothesis that there is no relation between χ and VRM. This hypothesis is strong in that it may be possible to specify rational structural models that yield reduced-form β_i for which χ and VRM are unrelated. Thus, if we reject the null hypothesis of no χ-VRM relation, we not only reject policy-invariant macromodels but also a subset of rational specifications as well.

---

[5] This third property is based on general observations on the length of standard business fluctuations and also reflects Friedman's claim that monetary shocks have their effects on real output over "long and variable lags" of 6 to 18 months.

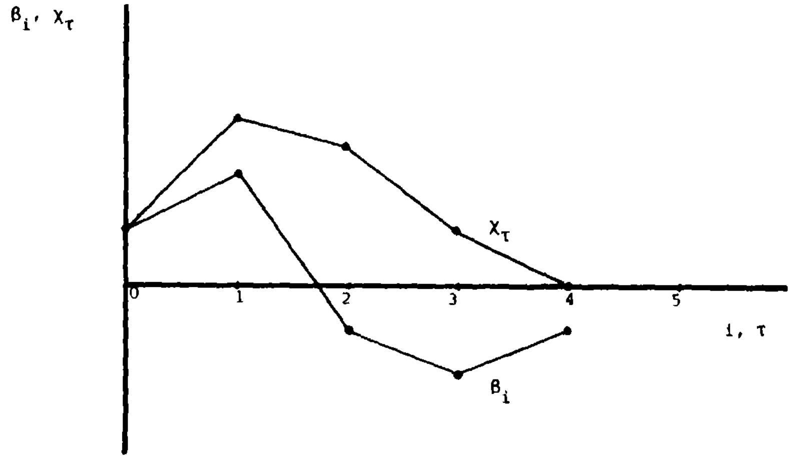


FIG. 1

III. The Methodology of the Tests

Our cross-regime tests of Lucas's propositions are based on the postwar annual data from 47 countries. The tests have three basic steps.[6] First, for each country a time series of the unanticipated component of money supply changes is estimated. Second, for each country a reduced-form real output equation is estimated. Third, using the estimates from steps 1 and 2, we test the hypothesis that χ and *VRM* are negatively related across countries. In this section, we describe the three steps in greater detail. Our approach in the first two steps is to develop procedures that can be applied uniformly to a large number of countries. We believe that such an approach is desirable because it allows us to test propositions about macroeconomic behavior without becoming mired in the idiosyncratic details of each country.

A. *The Data*

The data used in this study are from the International Monetary Fund's *International Financial Statistics* (*IFS*). The sample of 47 countries includes all countries for which the following annual series are available for a sufficiently long sample period:[7] (1) the money supply, M_t, defined as demand deposits and currency held by the nonbank public; (2) real income, Y_t, defined as gross domestic (national) product in 1975 prices. Further details are given in the Data Appendix.

---

[6] By combining the three steps into one estimation procedure one can gain efficiency Abel and Mishkin (1983), Pagan (1981), and Mishkin (1982) discuss this with respect to combining steps 1 and 2. Attfield and Duck (1982), in a paper closely related to this one, combine all three steps for an 11-country system and generate results similar to some in this paper. In general, using estimates from one regression as inputs into another such as we do in steps 2 and 3 can present econometric difficulties with respect to obtaining the correct test statistics. The paper by Pagan (1981) has a particularly useful discussion of some of these issues.

For the countries sampled, M_t and Y_t were nonstationary, and log-linear detrending did not seem to induce stationarity. For this reason, log first differences of M_t and Y_t, denoted as DM_t and DY_t, respectively, form the basis of our analysis. For a few countries DM_t was also nonstationary. For some countries, DM_t and/or DY_t exhibited some first- or second-order autoregressive structure or even some remaining nonstationarity in the mean. Table 1 presents some summary statistics for DM_t and DY_t.

B. Decomposing Money Supply Changes

The first step in our procedure is the decomposition of actual money supply changes, DM_t, into its unanticipated, RM_t, and anticipated, $DM_t - RM_t$, components. To generate this decomposition, assumptions must be made about the structure of the money supply process and the information available. Our assumptions are designed to yield a reasonable *estimate* of the unanticipated money supply changes and their variance by means of a parsimonious procedure that can be uniformly implemented for each country in the sample. Specifically, we assume that forecasts of money supply changes can be approximated by variants of the following equation[8]

$$DM_t = \alpha_0 + \alpha_1 DM_{t-1} + \alpha_2$$

A successful decomposition of DM_t
variant of (7) that has residuals, RM
noise.

With these procedures, we attempt
behavior in DM_t as reflected in its past

---

[7] The 47 countries in the sample are all those for which a real output series is available over at least 1957–77. Full details with respect to the data for each country are given in the notes to table 1, available from the authors on request.

[8] It has become clear from the work of Barro (1977, 1978), Barro and Rush (1980), Mishkin (1982), and others that at least for the United States the particular measure of unanticipated money supply used does not alter much the estimates of the coefficients of the real output effects of monetary shocks. Comparison of the results of Attfield and Duck (1982) with those in this paper confirms this for 10 other countries. In accordance with parsimony, the variant of (7) used suppresses α_i coefficients less than 1.5 standard errors from zero. This pretest procedure did not materially affect the estimated β coefficients or the estimated χ.

TABLE 1

Country	Years of Y (1)	No of DY (2)	μ_{DY} (3)	σ_{DY} (4)	Years of M (5)	No of DM (6)	μ_{DM} (7)	σ_{DM} (8)
Argentina	1950–77	27	.037	.040	1948–77	29	.341	.320
Australia	1949–77	28	.044	.024	1948–77	29	.065	.070
Austria	1949–77	28	.051	.027	1949–77	28	.091	.057
Belgium	1953–77	24	.039	.021	1948–77	29	.055	.042
Bolivia	1950–77	27	.034	.043	1950–77	27	.273	.260
Brazil	1952–77	25	.067	.063	1948–77	29	.305	.124
Burma	1950–77	27	.042	.051	1948–77	29	.096	.125
Canada	1948–77	29	.049	.025	1948–77	29	.056	.054
Chile	1950–77	27	.033	.042	1948–77	29	.488	.371
Colombia	1950–77	27	.050	.014	1948–77	29	.169	.061
Denmark	1950–77	27	.036	.026	1948–77	29	.074	.065
Dominican Republic	1950–77	27	.058	.056	1950–77	27	.090	.112
Ecuador	1950–77	27	.055	.035	1948–77	29	.132	.108
El Salvador	1951–77	27	.049	.020	1948–77	29	.087	.093
Finland	1950–77	27	.044	.033	1948–77	29	.099	.074
France	1950–77	27	.048	.016	1948–77	29	.113	.038
Greece	1950–77	27	.062	.034	1948–77	29	.165	.082
Guatemala	1950–77	27	.049	.022	1948–77	29	.086	.090
Honduras	1950–77	27	.039	.034	1948–77	29	.085	.090
Iceland	1948–77	29	.041	.050	1951–77	26	.168	.119
Ireland	1950–77	27	.030	.022	1948–77	29	.074	.051
Israel	1950–77	27	.082	.059	1948–77	29	.181	.089

Italy	1951–77	26	047	024	1948–77	29	134	.043
Japan	1952–77	25	081	.036	1953–77	24	.146	.058
Korea	1953–77	24	073	.038	1948–77	29	357	.229
Mexico	1948–77	29	060	.024	1948–77	29	.134	065
Netherlands	1948–77	29	042	027	1948–77	29	.071	057
New Zealand	1954–77	23	034	023	1948–77	29	.057	.076
Nicaragua	1950–77	27	057	038	1948–77	29	.104	.113
Norway	1950–77	27	.040	014	1948–77	29	.074	.051
Paraguay	1950–77	27	044	028	1948–77	29	196	.144
Peru	1950–77	27	050	040	1948–77	29	.163	.089
Philippines	1948–77	29	.059	018	1948–77	29	088	090
Portugal	1953–77	24	.052	031	1948–77	29	.085	.065
South Africa	1948–77	29	047	019	1948–77	29	.061	070
Spain	1954–77	23	.055	029	1950–77	27	.139	.051
Sri Lanka	1950–77	27	.040	031	1948–77	29	075	.100
Sweden	1950–77	27	032	020	1948–77	29	.069	.051
Switzerland	1949–77	28	036	032	1948–77	29	062	.044
Taiwan	1951–77	26	079	027	1950–77	27	.199	.100
Thailand	1957–77	20	073	019	1948–77	29	093	.066
Turkey	1948–77	[illegible]	059	039	1949–77	28	179	.069
United Kingdom	1948–77		026	019	1948–77	29	050	.058
United States	1948–77		035	027	1948–77	29	039	.025
Uruguay	1955–77		011	030	1948–77	29	.291	.209
Venezuela	1950–77		064	.026	1948–77	29	.113	106
West Germany	1950–77		[illegible]	030	1948–77	29	097	040

NOTE —*DY* and *DM* are the differenced natural l[illegible] stand for mean and standard deviation of the series *x*, *x* being either *DM* or *DY*. Full details of the data are in notes to this table, available from the authors on [illegible] information useful in forecasting DM_t may be available and because DM_t itself may be a noisy measure of the true DM_t, our RM_t may be an "errors-in-variables" measure of the true RM_t. The effects of this measurement problem are twofold. First, our measure of VRM may itself be an errors-in-variables measure of the true VRM. Second, use of RM_t as a regressor in an equation such as (5) may produce biased estimates of β_i. These two effects taken together can, under some conditions, lead to potential problems in testing hypotheses about the relation between β_i and VRM. The nature of these potential problems will be elaborated in Section III*E* below.

C. Step 2—the Real Output Equation

The second step in our procedure is the estimation of the real output effects of unanticipated money supply changes, based on an operational form of equation (5). The cross-correlations of RM_t and DY_t for the countries in our sample generally showed no discernible relation between RM_t and DY_t for lags greater than four. Likewise, the correlograms of (detrended) DY_t generally revealed no clear spikes at lags greater than two. Thus, we chose $N_1 = 4$ and $N_2 = 2$ as our maximum lag lengths and estimated variants of the following equation,[9]

$$DY_t = \alpha + \sum_{i=0}^{4} \beta_i RM_{t-i} + \sum_{i=1}^{2} \gamma_i DY_{t-i} + \delta t + RY_t, \qquad (8)$$

where β_i, γ_i, and δ are OLS projection coefficients and RY_t is the residual. In principle, one could test for properties (P1) through (P3) by estimating (8) both constrained and unconstrained and testing the null hypothesis that the constraints are not binding. In this paper we formally test (P2) but confine ourselves to an informal examination of (P1) and (P3).

D. Step 3—the Cross-Regime Tests

The third step in our procedure is to test the strong form of the rational expectations hypothesis, that is, a negative relation between χ and VRM, against the "nonrational" null hypothesis that there is no such relation. In structuring the cross-regime tests, we assume that for country j the true χ_j and VRM_j are generated as observations from the pair of random variables (χ, VRM) with a fixed joint distribution. Under the nonrational null hypothesis, χ and VRM are independently distributed, and under the strong form of the rational expectations hypothesis χ and VRM should be negatively related. In addition to tests of the relation between χ and VRM, we will explore related rational expectations propositions by examining the relation between χ and the variance of velocity shocks, the variance of real output shocks, and the mean and variance of actual money growth.

---

[9] In accordance with parsimony, the variant of (8) reported suppresses γ_i and δ coefficients less than 1.5 standard errors from zero. This pretest procedure did not materially affect the estimated β_i or the estimated χ. Also, experiments with lags of DY_t greater than two and lags of RM_t greater than four had little effect on the estimated β_i and χ.

E. Problems in Interpreting the Cross-Regime Tests

Our empirical implementation of cross-regime tests uses estimates of χ and VRM obtained from equations (7) and (8). The resulting estimation error in χ and VRM may affect the cross-regime relation. For example, in and of itself, estimation error in χ and VRM would tend to reduce the observed negative correlation between χ and VRM. In biasing inferences against the rational expectations proposition such effects are manageable. The possibility of an interaction of estimation error in (7) and estimation bias in (8), however, presents potentially more serious problems and hence bears more detailed discussion.

For the purposes of the ensuing discussion, we consider simple variants of (7) and (8). Assume first that for each regime DM_t is a white-noise process so that (7) reduces to $DM_t = RM_t$. Distinguish true RM_t^* from measured RM_t and assume that $RM_t = RM_t^* + v_t$, where v_t is measurement error and RM_t^* and v_t are orthogonal at all leads and lags. In this case $VRM = VRM^* + V_v$, where [illegible] true RM_t^* and V_v is the variance of v_t. As[illegible] the simplified form $DY_t = \Sigma_{i=0}^{N} \beta_i^* RM$[illegible] coefficients attached to true RM_t^* and [illegible] cludes any measurement error in DY_t, w[illegible] nal to RM_t^* and v_t at all leads and lags. [illegible] estimated consistently. Using RM_t data, [illegible] sistent estimates, β_i, with $\beta_i = \theta\beta_i^*$, where θ [illegible] Moreover, since θ is the same for all β_i, $\chi = \theta\chi^*$, where χ^* is the true "peak effect" as calculated from the true β_i^*, and χ is the measured "peak effect" as calculated from the estimated β_i.

In the case above, the relation between the observed χ and VRM and the true χ^* and VRM^* is given by the pair of equations

$$VRM = VRM^* + V_v \tag{9a}$$

and

$$\chi = \chi^*\left(\frac{VRM^*}{VRM^* + V_v}\right). \tag{9b}$$

Because of the V_v term in the two expressions, it may appear that a negative bias in the observed χ-*VRM* relation exists. In fact, if *VRM** is constant and the variation in observed *VRM* is due to V_v, a negative bias does arise. However, if V_v is constant across regimes and the variation in *VRM* is due to variation in true *VRM**, *VRM* and θ are positively correlated, thereby biasing one against finding a negative observed χ-*VRM* relation. Finally, if *VRM** and V_v vary in proportion across regimes, θ is a constant less than one, and the observed χ-*VRM* relation would be biased toward zero.

The considerations above suggest that the greater the cross-regime variation of the *VRM** relative to V_v and the more correlated is their variation, the more likely it is that the observed χ-*VRM* relation is biased *away from* a negative one. However, the possibility of a bias *toward* a negative χ-*VRM* relation does exist, thereby introducing problems in drawing inferences from an observed negative χ-*VRM* relation. Since the direction and magnitude of the bias are unknown, we present our cross-regime evidence in Section VI assuming no bias, but with the preceding discussion in mind. In addition, Monte Carlo results (available from the authors on request) show that in order for the negative bias to become severe enough to have a material effect on our cross-regime inferences, the observed cross-regime variation in *VRM* would have to be almost entirely attributable to cross-regime variation in V_v.[10]

IV. Decomposing Money Supply Changes—the Empirical Results

The results of estimating the unanticipated component of money supply changes in 47 countries, following the procedures set out in Section III*B*, are reported in table 2.[11] The second column presents the estimated model. The key diagnostic statistics used to determine the success of our decomposition describe the time-series properties of the residuals from the estimated models. The first-order through third-order autocorrelations and the summary statistic $Q(6)$ are in columns 7–10. The typical standard error of the autocorrelations is about .18. Thus, individual autocorrelations in the range of ±.25 and $Q(6)$ values less than about 9.0 are an indication of success. For most countries, the probability that the RM_t yielded by the parsimonious models in the second column are not white noise is low. For a few countries (e.g., Bolivia, Denmark, Netherlands) the specified models do not yield good diagnostics, usually because of one or two outlying observations. The chosen specifications were robust after these were deleted, but the reported RM_t series used all the data.

---

[10] Monte Carlo experiments are necessary because the expression for χ is too complicated to yield a general analytical expression for the bias in the observed covariance between measured χ and *VRM*. The experiments (using lognormal distributions since *VRM* appears lognormally distributed) reveal that when *VRM** and V_v have the same *cross-regime mean and variance, there is no bias and that when more (less) of the variation in VRM is due to VRM* than to* V_v, *a positive (negative) bias results in the observed* χ-*VRM* relation. The experiments also reveal that only if well over 90 percent of the observed cross-regime variation in *VRM* is due to V_v would the negative bias become severe enough to affect materially our inferences, given the strength of the empirical results reported in Sec. VI. For example, even if 90 percent of the observed variation in *VRM* were due to V_v, there would still be only about a 5 percent probability that the results were due to a spurious negative bias

[11] These results are discussed in greater detail in Kormendi and Meguire (1982).

V. The Real Output Equation—the Empirical Results

For each country in our sample we estimated variants of equation (8) according to the procedures set out in Section IIIC. The chosen variant of (8) was then estimated under the constraint that the sum of the β_i should equal zero, that is, that the monetary shocks should be neutral over the long run ([P2] in Sec. IIIC). As we shall see, this constraint was not rejected by the data for the set of countries in our sample. Thus, in table 3 we report the results of estimating (8) subject to the long-run neutrality constraint. The results for the unconstrained regressions are reported in an expanded version of table 3, available from the authors on request. For each regression, the diagnostic statistics reported include the first three residual autocorrelations (cols. 17–19) and the $Q(6)$ value (col. 20).
the autocorrelations are around .20 so t
of ±.30 and $Q(6)$ values less than abo
noise residuals.

All reported regressions were estim
of the available data, and the estimat
similarity. For five countries, we fail
satisfied this criterion.[12] These five pr
ported in table 3 after the other 42 countries
to variation in the sample period for each of these countries is severe enough to lead us to regard them with some suspicion. Thus, in what follows, the results we report are based on the set of 42 robust coun-

---

[12] The five countries were Austria, Greece, Mexico, Sweden, and Uruguay. The nonrobustness can be characterized as follows: deleting the first and or last five observations caused the estimates of the β_i and χ for these countries to change dramatically whereas for the other 42 countries the β_i and χ remained relatively stable. The results for these five countries reported in table 3 are basically "representative" (in that they lie in the middle of the range of estimates obtained) and employ as parsimonious a specification as possible.

TABLE 2

Country	RM Model (1)	Years of RM (2)	No. of RM (3)	(4)	σ_{DM} (5)	σ_{RM} (6)	r_1 (7)	r_2 (8)	r_3 (9)
Argentina	$RM = D2M$	1950–77	28	.320	.205	−.10	−.28	12	4.8
Australia	$RM = DM$	1949–77	29	.070	070	14	15	24	5.0
Austria	$RM = DM$	1950–77	28	.057	057	07	−.21	.28	7.1
Belgium	$DM = 019 + .003t + RM$	1949–77	29	.042	026	− 16	−.13	.22	8.8
	(1 8) (4 5)								
Bolivia	$RM = DM$	1951–77	27	.260	.260	55*	.27	24	14 0
Brazil	$RM = D2M$	1950–77	28	.124	.081	−.27	−.19	14	5.3
Burma	$RM = DM$	1949–77	29	.125	.125	−.01	−.11	− 02	4 2
Canada	$RM = DM$	1949–77	29	.054	.054	− 06	− 02	−.03	2.3
Chile	$RM = D2M$	1950–77	28	.371	179	05	20	−.04	9.1
Colombia	$RM = DM$	1949–77	29	.061	061	03	−.03	20	6.4
Denmark	$RM = DM$	1949–77	29	.065	.054	.45*	23	.28	19.1
Dominican Republic	$RM = DM$	1951–77	27	.112	112	− 17	.01	.23	4.2
Ecuador	$DM = .035 + 007t + RM$	1949–77	29	.108	093	− 33	.29	−.11	12.2
	(1.0) (3 1)								
El Salvador	$DM = .024 + .73DM\text{-}2 + RM$	1949–77	29	.093	079	11	−.12	.16	3.4
	(2 7) (6.1)								
Finland	$RM = DM$	1949–77	29	074	074	20	.15	.16	3.2
France	$DM = 06 + 42DM\text{-}1 + RM$	1950–77	28	.038	.028	.04	−.09	.10	10.8
	(3.0) (3.0)								
Greece	$RM = DM$	1949–77	29	.082	.082	−.09	.28	−.23	11.6
Guatemala	$DM = 048 + 37DM\text{-}1 + RM$	1949–77	29	.090	.079	13	.05	.19	5.8
	(4.0) (1 9)								
Honduras	$RM = DM$	1949–77	29	.090	090	.10	−.03	.25	2.7
Iceland	$DM = .07 + 007t + RM$	1952–77	26	.119	105	−.03	− 23	.18	6.0
	(1.6) (2 5)								
Ireland	$DM = 014 + .004t + RM$	1949–77	29	051	039	21	.12	08	3.2
	(1.6) (4.6)								
Israel	$RM = DM$	1949–77	29	.089	089	28	−.04	−.19	6.2
Italy	$DM = .10 + .002t + RM$	1949–77	29	.043	039	.19	−.05	.07	9.2
	(6.6) (2 8)								
Japan	$RM = DM$	1954–77	24	.058	.058	.18	14	.06	7.3
Korea	$RM = D2M$	1950–77	28	229	152	− 16	−.24	.15	6.0
Mexico	$RM = DM$	1949–77	29	.065	.065	.27	−.02	.07	4.9

Netherlands	$DM = 01 + 004t + RM$ (.6) (4 0)	1949–77	29	057	.046	−.04	−.39*	.07	9.3
New Zealand	$RM = DM$	1949–77	29	076	076	.36*	−.04	.11	6.2
Nicaragua	$DM = .07 + .36DM\text{-}1 + RM$ (2.8) (2.0)	1949–77	29	.113	.106	0	−.11	15	6.6
Norway	$DM = -010 + 005t + RM$ (−1.0) (9 8)	1949–77	29	.051	.023	.03	.07	−.09	3.4
Paraguay	$DM = 03 + 005t + 32D + RM$ (.4) (1.7) (5.5)	1949–77	29	144	089	17	13	.04	6.6
Peru	$RM = DM$	1949–77	29	.089	089	06	.11	.08	4.8
Philippines	$DM = -017 + 007t + RM$ (−.7) (4 6)	1949–77	29	090	.068	− 45*	.21	−.07	9.2
Portugal	$DM = 024 + 004t + RM$ (1.5) (3 8)	1949–77	29	.065	053	03	− 14	− 04	1.4
South Africa	$RM = DM$	1949–77	29	070	070	.24	.05	14	5.8
Spain	$DM = 077 + 45DM\text{-}1 + RM$ (6 4) (2 5)	1951–77	27	051	.046	07	−.07	−.21	1.8
Sri Lanka	$RM = DM$	1949–77	29	100	100	.24	−.33	−.16	11.6
Sweden	$RM = DM$	1949–77	29	051	051	− 08	− 09	.09	2.0
Switzerland	$RM = DM$	1949–77	29	044	044	.40*	11	−.07	11.1
Taiwan	$RM \quad DM$	1951–77	27	.100	100	19	08	.08	2.7
Thailand		1949–77	29	066	066	06	− 11	−.01	.9
Turkey	$DM = \quad + RM$	1950–77	28	.069	036	− 13	18	− 35*	7.1
United Kingdom	$DM =$	1949–77	29	058	044	− 06	− 04	−.15	4.6
United States	DM	1949–77	29	025	022	21	− 18	02	5.3
Uruguay	DM	1949–77	29	209	152	− 02	18	− 45*	11 4
Venezuela	DM	1949–77	29	106	088	16	− 05	11	1 8
West Germany		1949–77	29	040	.040	08	−.03	30	4.9

NOTE — RM is DM transformed to st
time trend. D for Paraguay and Turke
which was estimated OLS. The number
take into account the reduction in deg
including the first three reported obs
observations. The reported specification
available from the authors on request

model given in the second column. $DM\text{-}i$ denotes the ith lag of DM, $D2M$ is the difference of DM. The variable t is a
riable is dictated by inspection of the data. Models with $DM\text{-}i$ were estimated as ARMA models except for France,
ession coefficients are t statistics. σ_{DM} is from table 1. σ_{RM} is the standard deviation of RM and is computed so as to
parameters fit in estimating RM. $Q(6)$ is the Box-Pierce statistic adjusted for bias, computed using the first six r_i.
he 5 percent level are indicated by an asterisk. Most of the large r can be attributed to one or two outlying RM,
r's not sensitive to deleting the outlying observations. Full details of the data and results are in notes to this table

TABLE 3

	RM					*DY*													
COUNTRY	0 (1)	−1 (2)	−2 (3)	−3 (4)	−4 (5)	−1 (6)	−2 (7)	T (8)	χ (9)	N (10)	F_1 (11)	F_2 (12)	R^2 (13)	$\bar{R}^2$ (14)	SDR (15)	r_1 (16)	r_2 (17)	r_4 (18)	$Q(6)$ (19)
Argentina (1955–77)	−03 (−9)	01 (4)	03 (8)	02	−03 (−8)		−41 (−1 7)		01	23	44 (4,17) 78	2.52 (1,16) 13	33	13	00343	−12	0	14	3 2
Australia (1953–77)	11 (1 4)	15 (2 5)	−10 (−1 8)	−01	−15 (−2 0)		−42 (1 9)		26	25	2 87 (4,19) 05	1 99 (1,18) 18	41	26	00204	.21	08	19	5 8
Belgium (1954–77)	−.06 (−5)	44 (3 2)	−02	−20 (−1 3)	−16 (−1 1)				38	24	3 27 (4,19) 03	.16 (1,18) 69	.41	28	00180	26	15	−23	4.5
Bolivia (1955–77)	0 (2)	−05 (−1 6)	04 (1 3)	−06 (−1.8)	07			003 (3 3)	<0	23	1.91 (4,17) 16	4 96 (1,16) 04	51	.36	00275	−.08	−05	07	4 1
Brazil (1954–62, 1964–77)	.01 (1)	29 (2 6)	0	−16 (−1 5)	−14 (−1 3)			005 (3 3)	30	23	1 94 (4 17) 15	18 (1,16) 68	49	33	00458	.19	11	−.09	8 7
Burma (1953–77)	0 (−1)	−01 (−1)	07 (1 0)	−08 (−1 2)	02	−43 (−2 2)			06	25	.49 (4,19) 74	32 (1,18) 58	.28	09	00482	−.06	02	.18	6 8
Canada (1953–77)	−04 (−6)	25 (3 4)	−09 (−1 2)	−03 (−4)	−09				21	25	2 88 (4,20) .05	09 (1,19) 77	37	24	00222	07	−10	−19	2.4
Chile (1954–74, 1976)	−01 (−2)	03 (8)	−06 (−1 5)	−01 (−2)	05		−53 (−2 2)		02	22	1 10 (4,16) 39	1 61 (1 15) 22	40	.22	00277	34	−03	07	4 6
Colombia (1953–77)	09 (1 6)	.02 (4)	−02 (−5)	0	−09 (−1 8)	35 (1 7)			11	25	1 26 (4,19) 32	03 (1,18) .86	25	05	00139	0	21	−.22	3 9

Denmark (1953–77)	− 05 (− 4)	24 (1 8)	− 27 (− 2 4)	08 (6)					19	25	1.75 (4,20) 18	1 28 (1,19) 27	26	11	00242	05	−.24	14	8 3
Dominican Republic (1955–77)	03 (3)	17 (1 7)	− 01 (− 2)	− 20 (− 2 0)	01				20	23	1 38 (4,18) 28	32 (1 17) .58	24	07	00553	− 05	−.12	− 12	3 4
Ecuador (1953–72, 1974–77)	15 (2 4)	− 06 (− 1 2)	−.06 (− 1 3)	− 03 (− 5)	0				15	24	2 24 (4,19) 10	3 56 (1,18) 08	32	.18	00215	− 15	01	23	2 7
El Salvador (1953–77)	03 (7)	07 (1 3)	− 06	10 (1 7)	− 11 (− 2 7)				10	25	2 01 (4,20) .13	.36 (1 19) 56	29	14	00192	11	10	−.06	3 2
Finland (1953–77)	07 (1 1)	13 (1 6)	− 15 (1 5)	− 06 (− 5)	01	35 (1 6)	− 45 (− 2 2)		20	25	1 86 (4,18) 16	37 (1,17) 55	46	28	00275	− 04	− 02	− 13	6 4
France (1954–74, 1976–77)	08 (1 0)	14 (1 4)	− 12 (1 3)	− 09 (9)	01				22	23	1 58 (4,18) 22	50 (1,17) 49	.26	10	00119	− 31	09	.17	7.0
Guatemala (1953–77)	10 (1 9)	02 [illegible]	06	06	0				12	25	1 42 (4,20) 26	1 11 (1,19) 31	22	06	00198	0	0	− 23	4 5
Honduras (1953 1955–77)	05 (8)	[illegible]			02				12	24	62 (4,19) 65	1 37 (1 18) 26	12	0	00294	22	− 08	06	7 2
Iceland (1956–66 1969–77)	08 (8)				[illegible]				16	20	61 (1,15) 66	1.77 (1,14) 20	14	0	00398	30	− 09	− 10	2 8
Ireland (1953–77)	27 (3 5)					61 (3 6)			27	25	6 30 (4,19) 0	61 (1,18) 44	64	55	.00154	− 01	08	− 15	9 7
Israel (1953–77)	04 (4)								19	25	1 14 (1,20) 37	9 55 (1,19) 01	19	02	00460	20	− 03	12	2 3
Italy (1953–77)	10 (1 1)							002 (3 0)	.29	25	3 10 (4 19) 04	08 (1,18) 78	51	38	00191	− 13	− 16	− 33	5 8

TABLE 3 (*Continued*)

	RM					*DY*													
COUNTRY	0 (1)	−1 (2)	−2 (3)	−3 (4)	−4 (5)	−1 (6)	−2 (7)	T (8)	χ (9)	N (10)	F_1 (11)	F_2 (12)	R^2 (13)	$\bar{R}^2$ (14)	SDR (15)	r_1 (16)	r_2 (17)	r_3 (18)	$Q(6)$ (19)
Japan (1958–77)	12 (1 3)	20 (2 0)	− 37 (−3.3)	−.03 (− 3)	.08	55 (2 4)			32	20	4 29 (4,14) 02	06 (1,13) 81	61	48	.00284	0	− 12	.28	2.5
Korea (1955–77)	− 01 (− 2)	− 01 (− 2)	− 01 (− 2)	0	03 (.6)	− 33 (−1 4)		005 (3.8)	<0	23	11 (4,16) 98	85 (1,15) 37	.51	33	00318	− 09	− 02	.14	2 5
Netherlands (1953–77)	−.04 (− 5)	12 (1 3)	19 (2.3)	− 20 (−2 1)	− 07				27	25	3 76 (4,20) 02	1 92 (1,19) 18	43	32	.00210	− 02	−.04	− 03	1.4
New Zealand (1955–77)	08 (1 3)	07 (.9)	− 06 (− 8)	− 09 (−1 5)	0				15	23	1 86 (4,18) 16	50 (1,17) 49	29	.14	00217	31	− 08	− 16	6.6
Nicaragua (1953–77)	−.06 (−1.2)	16 (2 8)	− 06 (−1 1)	05 (.8)	− 09				10	25	1 93 (4,20) 14	2 90 (1,19) 10	28	13	.00310	−.04	.23	−.10	10 7
Norway (1953–77)	− 04 (− 3)	26 (2 6)	− 15 (−1 3)	− 13 (−1 3)	06				22	25	2 16 (4,20) .11	06 (1,19) .81	.30	16	00123	.14	− 10	05	4.8
Paraguay (1953–78)	09 (1 7)	− 08 (−1.5)	− 06 (−1 1)	− 02 (−.3)	07			002 (2 3)	09	25	1.38 (4,20) 28	4 27 (1,19) 05	37	.21	00233	03	− 02	15	1 4
Peru (1953–77)	.15 (1 8)	− 02 (− 2)	− 09 (−1.2)	01 (.2)	−.05				15	25	91 (4,20) .48	06 (1,19) .81	.15	0	.00371	.04	.04	−.25	2.9
Philippines (1953–77)	.08 (1 8)	.09 (1 9)	− 04 (−1 0)	− 13 (−3 1)	0		50 (3 2)		.17	25	2.72 (4,19) .06	29 (1,18) 60	45	31	.00128	− 11	.02	11	2.4

Portugal (1954–74, 1976–77)	12 (1 4)	− 08 (−.8)	− 11 (− 9)	05	02 (2)				12	23	1 08 (4,18) 40	.46 (1,17) 51	.19	01	.00234	−.08	0	.34	8.0
South Africa (1953–77)	.07 (1 3)	11 (1.8)	− 08 (− 1 3)	01	− 11 (− 2 4)	42 (2 0)			18	25	2 92 (4,19) 05	17 (1,18) 69	47	.33	00162	− 05	.11	04	5.5
Spain (1955–77)	.28 (2.8)	− 06 (− 6)	−.14 (− 1 4)	− 16 (− 1 6)	08				28	23	3 32 (4,18) 03	2 12 (1,17) 16	42	.30	00245	.23	− 05	08	4 1
Sri Lanka (1953–55, 1957–77)	− 06 (− 8)	0 (1)	15 (1 8)	− 16 (− 2 3)	07				09	24	1 65 (4,19) 20	16 (1,18) 69	.26	.10	00228	− 07	−.25	− 21	4.9
Switzerland (1953–77)	− 12 (− 1 3)	.29 (2 5)	.06 (4)	16 (1 3)	− 39			− 002 (− 2 5)	23	25	5 94 (4,19) 0	1 40 (1,18) 25	64	55	00213	23	− 31	− 14	6.8
Taiwan (1955–73, 1976–77)	03 (1 1)	07 (1 9)	− 06 (− 2 1)	01	− 05 (− 1 6)			002 (3 9)	10	21	2 55 (4,15) 08	51 (1,14) 49	74	66	00120	17	11	− 17	5 7
Thailand (1958–77)	11 (1 4)	02 (2)	− 08 (− 1 0)	11	15				13	20	1.41 (4,15) 28	80 (1,14) 39	27	08	00183	− 13	0	12	4 9
Turkey (1954–77)	16 (1 0)	− 07 (4)	[illegible]						16	24	1 46 (4,19) 25	68 (1 18) 42	24	07	00284	− 09	− 05	08	4 0
United Kingdom (1953–77)	01 (1)	17 (2 4)							18	25	1 71 (4 20) 19	04 (1,19) 84	25	11	00186	− 24	− 06	.29	5 3
United States (1953–77)	12 (7)	84 (1 1)							96	25	9 02 (4,20) 0	20 (1 19) 66	64	57	00166	03	− 03	− 24	2 0
Venezuela (1953–77)	06 (1 1)	− 05 (8)							09	25	63 (4,20) 65	19 (1,19) 67	11	0	00259	17	17	17	5 1
West Germany (1953–77)	06 (1)	24 (2 1)					28 (− 1 3)	004 (4 1)	26	25	1 53 (1,18) 24	16 (1,17) 69	60	47	00208	08	04	− 24	3 3

TABLE 3 (*Continued*)

Country	*RM* 0 (1)	 −1 (2)	 −2 (3)	 −3 (4)	 −4 (5)	*DY* −1 (6)	 −2 (7)	 T (8)	λ (9)	N (10)	F_1 (11)	F_2 (12)	R^2 (13)	$\bar{R}^2$ (14)	SDR (15)	r_1 (16)	r_2 (17)	r_3 (18)	$Q(6)$ (19)
Austria (1954–77)	−.18 (−2 1)	24 (2.6)	− 09 (− 1 2)	.04 (.5)	− 01				06	24	2 29 (4,19) 10	.29 (1,18) .60	.33	18	.00211	.27	18	−.08	5.5
Greece (1953–77)	02 (.2)	− 12 (−1.2)	0 (0)	0 (0)	.10 .				02	25	1.20 (4,20) .34	08 (1,19) .78	19	.03	00327	− 22	−.06	.22	5.7
Mexico (1955–77)	.09 (6)	.08 (.5)	08 (5)	− 20 (−1 4)	− 05 . .				25	23	.84 (4,18) .52	.06 (1,17) .81	16	0	00287	−.12	−.18	−.42*	7.4
Sweden (1953–77)	.09 (.9)	− 08 (− 9)	.06 (.7)	02	− 09 (−1 1)	62 (2 2)			09	25	.83 (4,19) .52	1.72 (1,18) 21	.25	.05	.00200	−.03	05	− 20	5.5
Uruguay (1958–77)	−.09 (−3 2)	06 (2 0)	.02 (8)	−.02 . .	03 (1.1)		− 57 (−2.9)	002 (2 4)	<0	20	2.65 (4,13) .08	20 (1,12) 66	70	56	.00207	−.01	−.04	−.23	4.6

NOTE.—The dependent variable is *DY*. Definitions, summary statistics, and diagnostics for *DY* and *RM* are given in tables 1 and 2. Estimates of the constant term are not reported. *t*-statistics are given in parentheses. The regressions are constrained by (P2) as discussed in the text and hence require one of the five lag coefficients to be the residual having no standard error, all that is at stake is the choice of the lag whose *t*-statistic *will not be defined*. *T* is a time trend. *N* is the sample size, the included years are listed under each country's name at the left. For some countries an outlier required deletion in order to achieve results robust to the sensitivity analysis discussed in Sec. V and n. 12. F_1 is the *F*-statistic for testing whether the *RM* lag coefficients are jointly zero. F_2 is the *F*-statistic for testing whether the zero-gain constraint is *not* binding. Under each *F*-statistic are listed its numerator and denominator degrees of freedom (in parentheses) and its marginal significance level. $\bar{R}^2$ is R^2 adjusted for degrees of freedom. SDR is the standard deviation of the residuals. r_i is the *i*th-order sample autocorrelation of the residuals. $Q(6)$ is the Box-Pierce statistic, adjusted for bias, computed over the first six r_i, including the first three reported here. Further details concerning the data and results are in notes to this table, available from the authors on request.

tries, but any differences that occur when all 47 countries are used are noted.

The first question about the β_i coefficients is whether they have significant incremental explanatory power. F-statistics, F_1, for the incremental significance of the set of β_i are given in column 11 of table 3, along with the relevant degrees of freedom and marginal significance levels. These reveal that almost half the countries reject at the 10 percent level the null hypothesis that the measured RM_{t-i} do not have explanatory power. A Kolmogorov-Smirnov (K-S) test of the marginal significance levels of the F-statistics rejects at the 1 percent level the null hypothesis that they are drawn from a [0, 1] uniform distribution and hence rejects the null hypothesis that the β_i over the set of 42 countries do not have explanatory power.

Next we ask whether leads of RM_t have significant incremental explanatory power when introduced into (8). Such tests are motivated by the work of Granger (1969) and Sims (1972) as tests of econometric exogeneity and as specification tests for equations such as (8). For each country we included two leads, RM_{t+1} and RM_{t+2}, into the output equations reported in table 3 and constructed an F-test for the null hypothesis that both coefficients are equal to zero. Only one out of 42 countries rejected the null hypothesis at the 10 percent level. A K-S test fails to reject at the 20 percent level the null hypothesis that the marginal significance levels of the set of 42 F-statistics are drawn from a [0, 1] uniform distribution.

Column 12 of table 3 presents the F-statistics, F, the relevant de-
grees of freedom, and the marginal signifi
whether long-run neutrality (P2) is
(Bolivia, Ecuador, Israel, and Nicar
levels are less than or equal to .10,
percent level in four out of 42 countr
ber under the null hypothesis. To te
of 42 countries is characterized by (P2
whether the marginal significance levels of
from a sample drawn from a [0, 1] uniform distribution. This test fails to reject the null hypothesis at the 20 percent confidence level.[14]

Given property (P2), property (P1) requires that any positive effects of RM_t on DY_t occur prior to the negative "reversals" that restore neutrality. Examination of the regression coefficients indicates that,

---

[13] Fisher (1984) investigates tests of the long-run neutrality of money and uncovers an interesting relation between such tests and tests of the *contemporaneous* exogeneity of real output with respect to money.

[14] For the highest variance countries, rational expectations models predict small real output effects of monetary shocks. Thus, we repeated the K-S test of (P2) after deleting the 15 lowest χ countries. The null hypothesis is still not rejected at the 20 percent level.

for most countries, this pattern holds reasonably well. We attribute the few deviations to sampling variability. The third property (P3) discussed in Section III simply states that the peak effect, χ, should occur prior to the second lag of RM_{t-i}. For all but three countries, either χ occurred prior to the second lag or the β_i that caused a violation of (P3) was insignificantly different from zero. Again, these results are consistent with the sampling variation to be expected across 42 countries. In light of these results, we take our measure of χ to be the maximum χ_τ for $\tau \leq 2$. We report this measure in column 10 of table 3.

We can abstract from sampling variation by examining the mean β_i coefficients across the 42 countries, denoting these as $\bar{\beta}_i$, $i = 0, 1, \ldots, 4$. Figure 2 plots the $\bar{\beta}_i$ and the associated $\bar{\chi}_\tau$.

Anticipating the cross-country tests of the next section, we rank the 42 countries by *VRM* and divide them into three groups. Group I contains the 15 countries with the smallest *VRM* values; group II, the middle 17; and group III, the largest 10. (These groups emerge naturally from an inspection of the empirical distribution of *VRM*.) The mean χ_τ for each group, $\bar{\chi}_{\tau,i}$, $i =$ I, II, III, are plotted in figure 3. The average peak effect for each group, $\bar{\chi}_i$, $i =$ I, II, III, is labeled on the vertical axis. At all lags, the effect of RM_t on DY is smaller for group III than for group II and likewise for groups II and I. The data plotted in figure 3 are interesting in two respects. First, they provide preliminary evidence for the stronger form of the Lucas hypothesis, that is, that χ is negatively related to *VRM*. Second, they show that our scalar measure χ should be an adequate summary representation of the β vector when testing the hypothesis.

VI. The Cross-Regime Evidence

A. *The Cross-Regime Evidence for Lucas's Proposition*

The strong form of Lucas's rational expectations proposition predicts a negative relation between the peak effects of monetary shocks on real output, χ, and the variance of those shocks, *VRM*. In this section we examine the evidence for this proposition over the countries in our sample. In the next section we examine some related propositions. We take our measure of χ from table 3, column 10, and our measure of *VRM* from table 2, column 5 (with $VRM = \sigma^2_{RM}$). These numbers are presented in table 4 for convenience. The main results we report are based on the set of 42 robust countries. Since the five nonrobust countries all had β_i coefficients that changed dramatically when small subsets of the data were deleted, we are suspicious of the resulting χ for these countries. As we shall see, however, the pooled results are similar to those of the 42 robust countries.

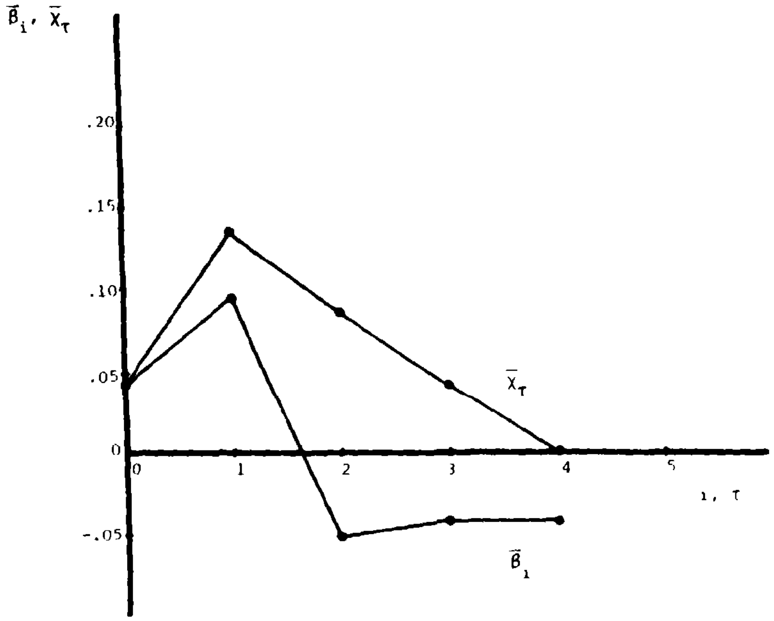

$\bar{\beta}_i, \bar{X}_\tau$
.20
.15
.10
.05
0
-.05
0
1
2
3
4
5
i, τ
$\bar{X}_\tau$
$\bar{\beta}_i$

FIG. 2

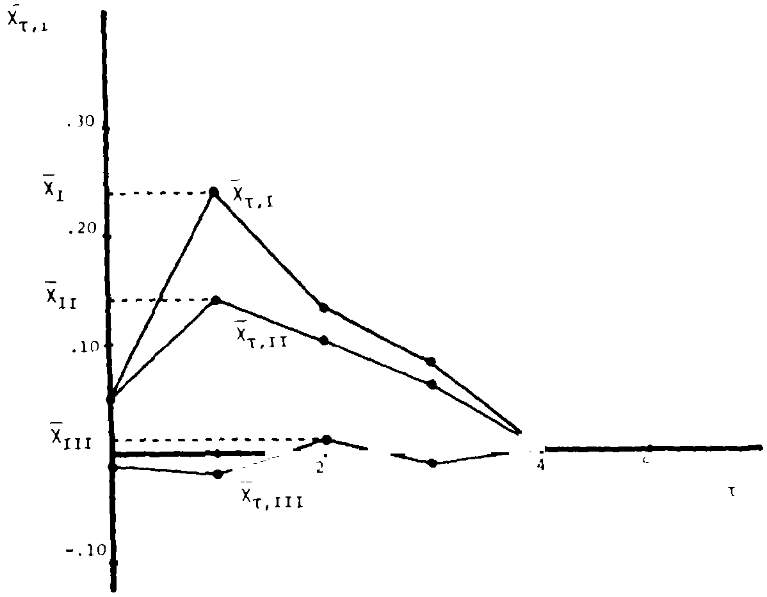

$\bar{X}_{\tau,i}$
.30
$\bar{X}_I$
.20
$\bar{X}_{II}$
.10
$\bar{X}_{III}$
-.10
$\bar{X}_{\tau,I}$
$\bar{X}_{\tau,II}$
$\bar{X}_{\tau,III}$
2
4
τ

FIG. 3

TABLE 4

	χ (1)	*VRM* (2)	*MDM* (3)	*VDM* (4)	*VDV* (5)	*VRY* (6)
Argentina	01	.0420	.341	.1024	.0566	00118
Australia	26	0049	065	.0049	0024	00042
Belgium	38	0007	055	0018	.0014	00032
Bolivia	.01*	0676	.273	.0676	0185	00076
Brazil	30	.0066	305	0154	.0053	.00210
Burma	06	0156	.096	.0156	0180	00232
Canada	21	0029	056	0029	.0046	00049
Chile	02	0320	488	1376	0317	.00077
Colombia	.11	.0037	.169	0037	0036	00019
Denmark	.19	0029	.074	.0042	0030	00059
Dominican Republic	.20	.0125	.090	0125	0144	.00306
Ecuador	.15	0086	.138	0117	0044	00046
El Salvador	10	.0062	087	0087	.0052	00037
Finland	20	0055	099	0055	0050	00076
France	22	0008	113	0014	.0017	00014
Guatemala	12	0062	.860	.0081	.0038	00039
Honduras	12	0081	085	.0081	.0077	.00086
Iceland	.16	.0110	168	0112	.0046	00158
Ireland	27	0015	074	.0026	0019	00024
Israel	19	.0079	181	0079	0074	00212
Italy	29	0015	.134	.0019	0013	00036
Japan	.32	.0034	116	0034	0045	00081
Korea	01*	0231	357	.0524	0207	00101
Netherlands	27	0021	.071	0033	0044	.00044
New Zealand	15	0058	057	0058	0027	00047
Nicaragua	.10	0112	104	0128	0098	00096
Norway	22	0005	074	0026	.0016	00015
Paraguay	09	0079	196	0207	.0098	00054
Peru	15	0079	.163	.0079	0048	00138
Philippines	17	0046	088	0081	0066	.00016
Portugal	12	0028	085	0042	.0030	00055
South Africa	18	0049	061	.0049	0020	00026
Spain	28	0021	139	0026	0013	00060
Sri Lanka	09	0100	075	0100	0071	00052
Switzerland	23	0019	062	0019	0024	00045
Taiwan	.10	0100	199	0100	.0072	.00014
Thailand	13	0032	093	.0014	0041	00033
Turkey	.16	.0013	179	0048	0033	00081
United Kingdom	18	0019	050	.0034	0023	00035
United States	.96	0005	039	0006	.0008	.00028
Venezuela	.09	.0077	.113	.0124	0046	.00067
West Germany	26	0016	097	0016	0029	00043
Austria	06	0033	091	0033	0036	00045
Greece	02	.0067	165	0067	0035	00107
Mexico	25	0042	134	0042	0041	00082
Sweden	09	.0026	069	0026	0012	.00040
Uruguay	01*	.0231	291	0437	.0289	00043

NOTE χ is taken from table 3, col 10 The asterisk means that χ was set equal to 01 to enable calculation of ln χ *VRM* is the square of σ_{RM} reported in table 2 *VDM* is the square of σ_{DM} reported in table 1 *VDV* is calculated as the variance of $\Delta(X_t/M_t)$, where X_t and M_t are defined in the Data Appendix *VRY* is the square of SDR reported in table 3

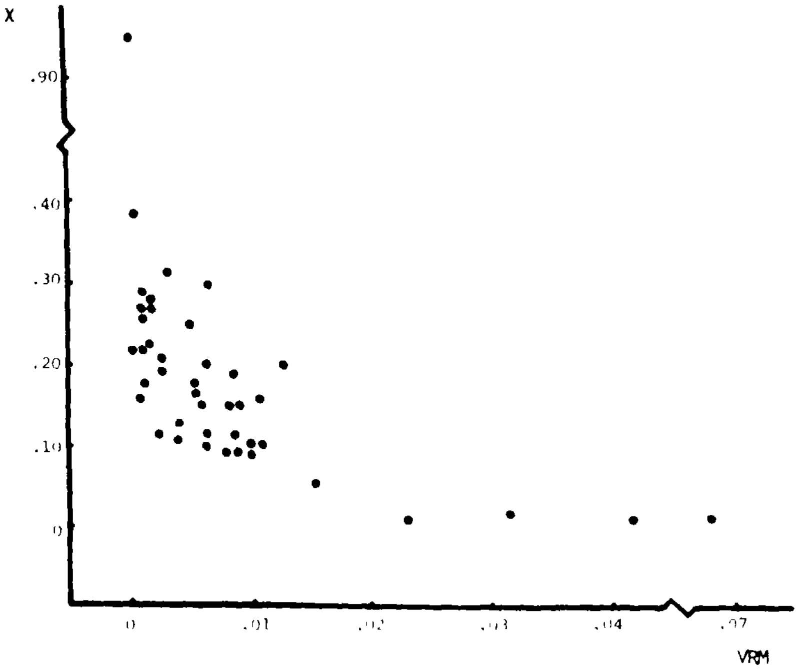


FIG. 4.—Scatter of χ and *VRM*

Figure 4 presents the data on χ_j and VRM_j for $j = 1, \ldots, 12$ in the form of a scatter diagram. Note the prono
ently nonlinear relation. Since the
relation suggested by rational exp
(1973) and Barro (1976) is itself non
by a semilog relation,[15] we present
ln χ_j and VRM_j. Herein lies the ess

---

[15] In Barro (1976) the analogue to χ is the coefficient on the monetary shock term in his eq. (15) on p. 11, which, in his notation, is

$$\left(\frac{H}{\alpha}\right)\left(\frac{\sigma_v^2 + \sigma_\epsilon^2}{\beta^2\sigma_m^2 + \sigma_v^2 + \sigma_\epsilon^2}\right)$$

Letting $\gamma_0 = (H/\alpha)$ and $\gamma_1 = \beta^2/(\sigma_v^2 + \sigma_\epsilon^2)$, we can write this coefficient, as $\chi = \gamma_0(1 + \gamma_1 VRM)$, where in a given regime γ_0 and γ_1 are positive constants that depend upon the parameters of the system, and $VRM = \sigma_m^2$. Taking the log of χ yields $\ln \chi = \ln \gamma_0 - \ln(1 + \gamma_1 VRM) \approx k_0 + k_1 VRM$, where $k_1 \leq 0$, and the linear approximation is better the smaller is $\gamma_1 VRM$. (Using a higher order approximation does not materially affect any of the empirical results in this paper.) Note that there may be cross-regime variation in the parameters γ_0 and γ_1. Thus, in our cross-regime tests, we are implicitly assuming that the cross-regime variation in γ_0 and γ_1 is independent of *VRM*.

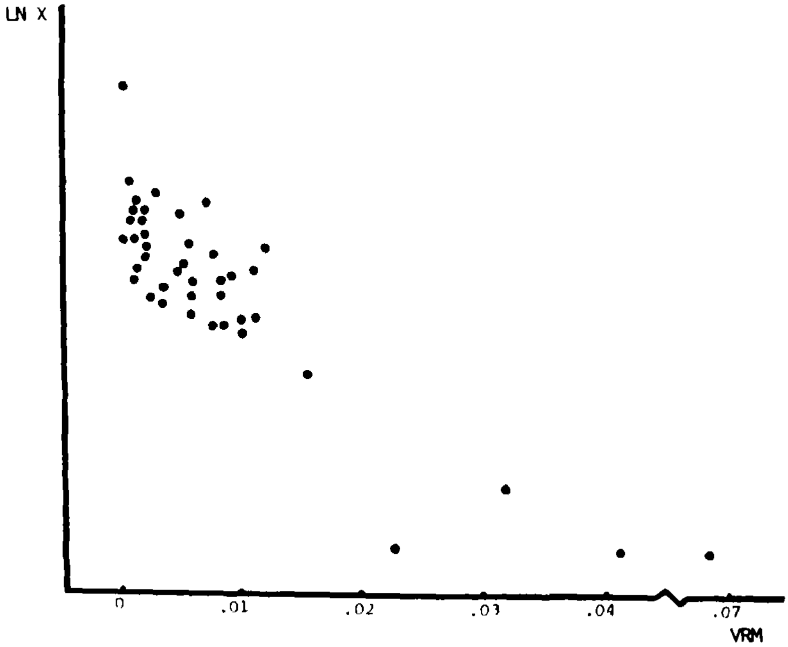


FIG. 5.—Scatter of ln χ and *VRM*

Lucas's hypothesis and against the nonrational null hypothesis. The semilog transformation produces a clearly negative and reasonably linear relation. Moreover, although the four extreme values of VRM_j enhance the negative relation, a clear negative relation exists for the main cluster of countries as well. This latter observation is particularly significant in light of the importance of the extreme observations for Lucas's (1973) empirical results.

As a means of calibrating the negative relation uncovered between χ_j and VRM_j and conducting some sensitivity analysis, we present in the first row of table 5 the correlation coefficients between ln χ_j and VRM_j for several subsets of the data. Since the data on χ_j and VRM_j were themselves generated through estimation procedures, their distributional properties are not fully known (see n. 5). Although this rules out precise hypothesis testing, we present the standard *t*-statistics in parentheses below the correlation coefficients and warn the reader to use them primarily as a heuristic for interpreting and comparing the strength of the observed correlations. We also employ nonparametric methods of evaluating the χ-*VRM* relation. The rank correlation coefficient provides a nonparametric basis for evaluating the χ-*VRM* relation that is also robust to the functional form. Thus, in the second row of table 5 we present the corresponding rank correlation coefficients (and their *t*-statistics in parentheses).

Column 1 of table 5 contains the ordinary and rank correlations for the basic set of 42 robust countries. The main point of interest here is that the nonparametric rank correlation provides almost as strong evidence for a negative χ-*VRM* relation as does the ordinary correlation. In light of Lucas's finding that his results turned on two extreme observations in his sample, we eliminate the five extreme observations in our sample—the highest χ country (the United States) and the four highest *VRM* countries (Argentina, Bolivia, Chile, and Korea)—and present the ordinary and rank correlations for the remaining 37 countries in column 2 of table 5. The main point of interest here is that although the negative relation diminishes somewhat, it remains quite strong. Finally, to illustrate the effects of adding the five nonrobust countries to the sample, we present the ordinary and rank correlations for the pooled set of 47 countries and for the five nonrobust countries in columns 3 and 4 of table 5, respectively. (Because of the small number of observations we do not report the *t*-statistics in col. 4.) Note that although the negative correlations for the pooled data are slightly smaller than for the 42 robust countries, the five nonrobust countries show strong negative correlations as well.[16]

B. The Cross-Regime Evidence for Related Rational Expectations Propositions

Rational expectations models predict that the [illegible]
monetary shocks should be related to [illegible]

TABLE 5

	42 Robust Countries (1)	37 Countries Excluding Extreme Observations (2)	[illegible]	[illegible]
Correlation	−.84	−.67	.78	−.82
	(−9.8)	(−5.3)	(8.4)	( )
Rank correlation	−.76	−.62	−.73	−1.00
	(−7.4)	(−4.7)	(−7.2)	(−)

NOTE.—We report the standard *t*-statistics in parentheses under the caveat that (for the ordinary correlation in particular) the distributional properties are unknown (see discussion in text). In col. 4 we do not report *t*-statistics because of the small number of observations.

---

[16] Although the negative correlation for the sample of five countries is almost as large as for the sample of 42 countries, there is considerable heteroscedasticity between the two samples. This is consistent with our finding that the χ's for the sample of five were not robust (see n. 12).

monetary shocks, *VRM*, but not to either the variance of *actual* monetary changes, *VDM*, or the mean growth of the money supply, *MDM*.[17] These predictions can be tested by examining the *partial* correlations between χ and the set of variables *VRM*, *VDM*, and *MDM*. Under the rational expectations hypotheses, the partial correlation between χ and *VRM*, with *VDM* and *MDM* held constant, $\rho(\chi, VRM \mid VDM, MDM)$, should be negative, but the other partial correlations, $\rho(\chi, VDM \mid VRM, MDM)$ and $\rho(\chi, MDM \mid VRM, VDM)$, should be zero.[18]

Since standard policy-invariant models yield the same predictions for the latter two partial correlations such evidence does *not* distinguish between rational and nonrational models. Such evidence is important, however, because our measures of *VRM*, *MDM*, and *VDM* are all positively correlated—$\rho(VRM, MDM) = .66$, $\rho(VRM, VDM) = .63$, and $\rho(MDM, VDM) = .79$—in which case the negative simple correlation between χ and *VRM* observed in table 5 may be the result of a failure to control for either *MDM* or *VDM*. The partial correlation between χ and *VDM* is of further interest, because such a zero partial correlation would mean that the variance of *anticipated* money supply changes, *VDM*-*VRM*, is uncorrelated with χ.

In table 4, we present the data on *VDM* and *MDM*. In table 6, column 1, we present the partial correlation coefficients for χ and *VRM*, χ and *VDM*, and χ and *MDM*, respectively, for the 42 robust countries. In column 2, we present the same partial correlations over the set of 37 countries that excludes the extreme observations. As with table 5, we present the *t*-statistics for the partial correlations primarily as a heuristic guide for interpreting and comparing the strength of the correlations.[19] There are several things to note. First, there is still a strong negative partial correlation between χ and *VRM*, even controlling for *VDM* and *MDM*. Second, although there is some evidence of a negative partial correlation between χ and *MDM*, it virtually disappears when the extreme observations in the data are omitted.[20] Third, there is virtually no correlation between χ and *VDM*

---

[17] In Lucas (1973), e.g., the two high variance of inflation countries are also the two high mean inflation countries. Thus, his results cannot distinguish these two potential factors.

[18] We obtain the partial correlation between χ and *VRM*, holding *MDM* and *VRM* constant, as the ordinary correlation between (1) the residuals of the orthogonal (least-squares) projection of χ onto *MDM* and *VDM* and (2) the residuals of the orthogonal projection of *VRM* onto *VDM* and *MDM*.

[19] The rank correlations corresponding to the partial correlations in table 4 are very similar.

[20] In fact, the evidence for a negative χ-*MDM* partial correlation can be traced to a single observation (Korea). Upon its deletion, the partial correlation between χ and *MDM* is .01.

TABLE 6

Partial Correlation between χ and:	42 Robust Countries (1)	37 Countries Excluding Extreme Observations (2)
VRM	−.68	−.35
	(−5.9)	(−2.2)
MDM	−.24	.16
	(−1.5)	(1.1)
VDM	−.06	−.10
	(−.4)	(−.6)

NOTE.—The partial correlation coefficient between χ and *VRM*, holding *MDM* and *VDM* constant, is calculated as the ordinary correlation between (1) and residuals from the orthogonal projection of χ onto *MDM* and *VDM* and (2) the residuals from the orthogonal projection of *VRM* onto *MDM* and *VDM*. We report the standard *t*-statistics under the caveat that the distributional properties are not fully known.

once *VRM* is held constant. This latter result is important, because it shows that while there is evidence of a strong negative relation between χ and the variance of *unanticipated* monetary shocks, *VRM*, there is no evidence of a relation between χ and the variance of *anticipated* money supply changes, *VDM-VRM*.

Rational expectations macromodels also generate implications concerning the effects of velocity shocks and aggregate real shocks on χ that can be examined using our data. By treating them isomorphically to monetary shocks, velocity shocks can be introduced into rational expectations models in a straightforward manner, in which case χ will depend negatively on the variability of both sources of aggregate nominal disturbances. However, in Barro (1980), because of asymmetric interactions with interest rates, the analogy between monetary and velocity shocks does not hold, and the predicted [illegible] χ and the variance of velocity shocks is am[illegible] the variability of velocity shocks we [illegible] changes in (log) velocity (measured as [illegible] the money supply), reported in table [illegible]

Aggregate supply shocks also have [illegible] expectations macromodels. In Barro ([illegible] ing aggregate supply shocks depend on [illegible] stochastic processes for monetary shocks, aggre[illegible] relative real shocks, and hence the predicted effect of the variance of aggregate real shocks on χ is ambiguous (see also Grossman and Weiss 1982). However, if real shocks are part of the noise that inhibits extraction of the monetary signal, the effects should be positive. Under the maintained assumption that aggregate real disturbances and aggregate nominal disturbances are orthogonal, the variance of the residuals of the real output regressions, *VRY*, reported in column 16 of table 3 and again in table 4, can be taken as a measure of the variability of aggregate real shocks.

TABLE 7

Partial Correlation between χ and:	42 Robust Countries (1)	37 Countries Excluding Extreme Observations (2)
VRM	-.62	- 60
	(-4.9)	(4.5)
VDV	- 48	-.30
	(-3.4)	(-1.9)
VRY	26	64
	(1 6)	(4.8)

Note. See notes to table 6

To address these two issues we present in table 7, column 1, the partial correlations for χ and *VRM*, χ and *VDV*, and χ and *VRY*, respectively, over the 42 robust countries. In column 2 we present the same partial correlation coefficients over the set of 37 countries that exclude the five extreme observations.[21] We also present in figures 6*a*, *b*, and *c* the scatter plots corresponding to the partial correlations in table 7 for the set of 37 countries that exclude the extreme observations. There are several things to note. First, the χ-*VRM* partial correlations both over the full set of 42 countries and after deleting the five extreme observations are still strongly negative. Second, both over the full set of 42 countries and after deleting the five extreme observations, the χ-*VDV* partial correlations are negative,[22] which is consistent with velocity shocks having effects similar to monetary shocks. Third, both over the full set of 42 countries and after deleting the extreme observations the χ-*VRY* partial correlation is positive, and in the latter case it is quite large.[23] Such results are consistent with output shocks functioning as noise in the extraction of monetary signals.[24] Although the rational expectations predictions are ambiguous, the evidence above is of interest because only rational expectations models among well-known macromodels give any reason to expect any effects of *VRY* and *VDV* on ln χ. Furthermore, these results provide an empirical basis for future modeling of real and velocity shocks.

---

[21] The rank correlations corresponding to the partial correlations in table 7 are very similar.

[22] The large negative partial correlation for the set of 42 countries is dominated by the five extreme observations.

[23] The smaller positive partial correlation between χ and *VRY* over the set of 42 countries is due to one outlying observation—the United States Upon its deletion the partial correlation between χ and *VRY* is .56.

[24] The positive χ-*VRY* relation might also emerge if the variance of aggregate real shocks and the variance of relative shocks were positively correlated across countries. See Kretzmer (1983) for evidence across industries in the United States that the variance of relative shocks is positively related to the output effects of monetary shocks.

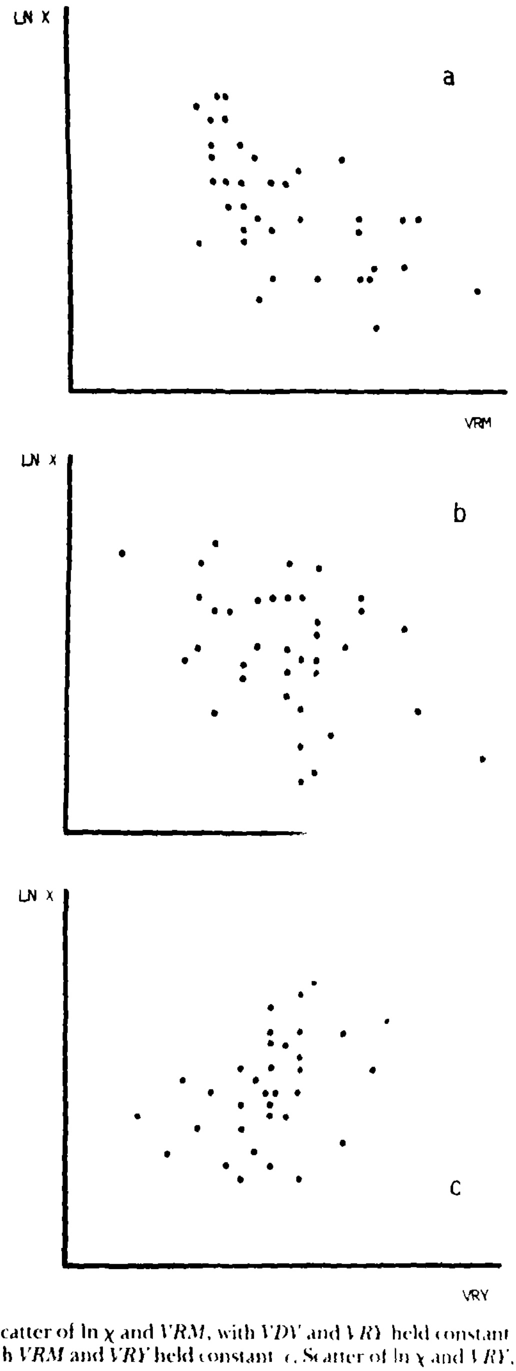


FIG. 6.—*a*, Scatter of ln χ and *VRM*, with *VDV* and *VRY* held constant. *b*, Scatter of ln χ and *VDV*, with *VRM* and *VRY* held constant. *c*, Scatter of ln χ and *VRY*, with *VRM* and *VDV* held constant.

VII. Conclusion

Much attention has been devoted recently to the development of rational expectations macromodels and to tests of propositions derived from them. Most of the tests to date have been undertaken in the context of a single policy regime. These tests have taken one of the following forms: (1) cross-equation restrictions implied by the rational expectations assumption (e.g., Sargent 1976*a*, 1978; Mishkin 1982); or (2) the neutrality of the anticipated components of certain policy variables with the nonneutrality of the anticipated component of those variables (e.g., Barro 1976, 1980; Mishkin 1982). Such single-regime tests are often difficult to execute or inconclusive because the identification problems are potentially severe (see Sargent 1976*b*). Very precise structural modeling is required for successful implementation of these tests, precision that is difficult to attain given the current state of theory and data availability.

In this paper we have taken an alternative approach to testing propositions derived from rational expectations models—the cross-regime approach first attempted by Lucas (1973). We have utilized the data from 47 countries representing as many different policy regimes. For each country, we have employed a variant of the same simple two-equation macromodel. The lack of structural precision in this two-equation model, although usually fatal to single-regime tests, is manageable in the context of our cross-regime tests.

There are many findings in this paper; we summarize some of the more important ones. First, for the countries in our sample, the lagged effects of monetary shocks on changes in real output have the following reasonably consistent characteristics: (1) current period monetary shocks have generally positive effects contemporaneously and one period lagged; (2) those positive effects are generally reversed through the second, third, and fourth lags, so that long-run neutrality of monetary shocks with respect to real output changes characterizes the set of countries in our sample; and (3) for most countries, the peak effect of a current period monetary shock on real output generally occurs within a year of the shock.

Second, the following implications of rational expectations models with respect to the determinants of the size of the real output effect of a monetary shock (χ) characterize the set of countries in our sample: (1) There is a significant and robust negative relation between χ and the variance of monetary shocks, *VRM*. This is consistent with Lucas's basic rational expectations proposition. Moreover, it rejects stable Phillips curve–type structural models, whereas Lucas's earlier test, properly interpreted, cannot do so. (2) There is evidence of a negative relation between χ and the variance of velocity changes, *VDV*. (3)

There is evidence of a positive relation between χ and the variance of aggregate real shocks, *VRY*. (4) Controlling for *VRM*, there is no discernible relation between χ and either the variance of actual money supply changes or the mean growth rate of the money supply.

Data Appendix

All data for a given country, except price indices, are in units of own currency.

Y is real gross domestic product (*IFS* series 99b.p or 99b.r) or real gross national product (*IFS* series 99a.p or 99a.r), depending on availability, and measured in 1975 prices. Pre-1960 data for Austria and the Netherlands are from *European Historical Statistics* (Mitchell 1974). For Brazil, Mexico, and Peru, *Y* is GDP in current prices (*IFS* series 99b) divided by an index of consumer prices (*IFS* series 64) for the last month of the year, with 1975 as the base year.

X, output in current prices, is either GDP from above or GNP (*IFS* series 99a), whichever is comparable to the available *Y* series for a country. Pre-1958 values for Nicaragua and pre-1950 values for all other countries were obtained from 1978 and 1979 versions of the *IFS* tape, respectively.

M (*IFS* series 34), end-of-period money, includes (*a*) currency outside banks and the government treasury and (*b*) demand deposits held by residents of the country in its commercial banks, the central bank, or the postal checking system (if any). Additional significant digits were taken from the 1971 *IFS Yearbook* for Argentina, Chile, and Uruguay. For Paraguay, *M* is currency outside banks (*IFS* series 14a) plus demand deposits at commercial banks (*IFS* series 24).

Further details on the data are in the notes to tables 1 and 2, available from the authors on request.

References

Abel, Andrew B., and Mishkin, Frederic S
Rationality, Market Efficiency and the
Policy." *J. Monetary Econ* 11 (January
Alberro, José. "The Lucas Hypothesis on the
tional Evidence." *J Monetary Econ.* 7 (Ma
Attfield, Clifford L. F., and Duck, Nigel W
Money Growth on Real Output." Working
1982.
Barro, Robert J. "Rational Expectations and the Role of Monetary Policy." *J Monetary Econ.* 2 (January 1976): 1–32.
———. "Unanticipated Money Growth and Unemployment in the United States." *A.E.R.* 67 (March 1977): 101–15.
———. "Unanticipated Money, Output, and the Price Level in the United States." *J.P.E.* 86 (August 1978): 549–80.
———. "A Capital Market in an Equilibrium Business Cycle Model." *Econometrica* 48 (September 1980): 1393–1417.
———. "The Equilibrium Approach to Business Cycles." In *Money, Expectations, and Business Cycles: Essays in Macroeconomics.* New York: Academic Press, 1981.

Barro, Robert J., and Rush, Mark. "Unanticipated Money and Economic Activity." In *Rational Expectations and Monetary Policy,* edited by Stanley Fischer. Chicago: Univ. Chicago Press (for N.B.E.R.), 1980.

Fisher, Mark E. "Long-Run Neutrality and the Exogeneity of Money." Working Paper. North Carolina State Univ., April 1984.

Froyen, Richard T., and Waud, Roger N. "Further International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 70 (June 1980): 409–21.

Granger, C. W. J. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37 (July 1969): 424–38.

Grossman, Sanford J., and Weiss, Laurence. "Heterogeneous Information and the Theory of the Business Cycle." *J.P.E.* 90 (August 1982): 699–727.

International Monetary Fund. *International Financial Statistics,* various monthly issues.

———. *International Financial Statistics, Annual Supplement.* Washington: IMF, 1971, 1972.

———. *International Financial Statistics Yearbook.* Washington: IMF, 1979, 1980.

Kormendi, Roger C., and Meguire, Phillip G. "Cross-Regime Tests of Macroeconomic Rationality." Working Paper. Univ. Chicago, October 1982.

Kretzmer, Peter E. "Cross-Industry Tests of an Equilibrium Business Cycle Model with Rational Expectations." Working Paper. Univ. Chicago, October 1984.

Lucas, Robert E., Jr. "Expectations and the Neutrality of Money." *J. Econ. Theory* 4 (April 1972): 103–24.

———. "Some International Evidence on Output-Inflation Tradeoffs." *A.E.R.* 63 (June 1973): 326–34. "Errata." *A.E.R.* 66 (December 1976): 985.

———. "An Equilibrium Model of the Business Cycle." *J.P.E.* 83 (December 1975): 1113–44.

McCallum, Bennett T. "On the Observational Inequivalence of Classical and Keynesian Models." *J.P.E.* 87 (April 1979): 395–402.

Mishkin, Frederic S. "Does Anticipated Monetary Policy Matter? An Econometric Investigation." *J.P.E.* 90 (February 1982): 22–51.

Mitchell, Brian R. *European Historical Statistics, 1750–1970.* New York: Columbia Univ. Press, 1975.

Pagan, Adrian. "Econometric Issues in the Analysis of Regressions with Generated Regressors." Working Paper. Australian National Univ., November 1981.

Sargent, Thomas J. "A Classical Macroeconometric Model for the United States." *J.P.E.* 84 (April 1976): 207–37 (*a*)

———. "The Observational Equivalence of Natural and Unnatural Rate Theories of Macroeconomics." *J.P.E.* 84 (June 1976): 631–40. (*b*)

———. "Estimation of Dynamic Labor Demand Schedules under Rational Expectations." *J.P.E* 86 (October 1978) 1009–44.

——— *Macroeconomic Theory.* New York: Academic Press, 1979.

Sims, Christopher A. "Money, Income, and Causality." *A.E.R.* 62 (September 1972): 540–52.

Weber, Warren E. "Prior Information and the Observational Equivalence Problem." *J.P.E.* 89 (April 1981): 411–15.

On Success and Dissolution of the Labor-managed Firm in the Capitalist Economy

---

Hajime Miyazaki

Stanford University

Labor-managed firms (LMFs) in the developed capitalist economy exhibit an interesting life-cycle pattern. Often they come into existence as a way of reorganizing insolvent entrepreneurial (profit-maximizing) firms. Yet successful LMFs tend to be assimilated to the entrepreneurial modus operandi in the long run. I propose a simple theoretical model that explains this LMF phenomenon as the outcome of the LMF's interaction with outside markets, especially competitive spot labor markets. The interpretations and limitations of the model are also discussed as the basic framework is extended to deal with imperfections in capital markets and with [illegible] markets.

I. Introduction

I consider here the establishment and [illegible] firms (LMFs) within the context of a deve[illegible]omy.[1] Specifically, I try to give a simple theore[illegible] following two points. First, the LMF is frequently established as a way

---

I thank Avner Ben-Ner for valuable discussions. In writing the present version, I have benefited from comments of Masahiko Aoki, Paul Evans, Itsuko Funayama, James Heckman, Thomas MaCurdy, Hugh Neary, Jan Svejnar, David Weisberg, and Martin Weitzman. I am solely responsible for any remaining errors and ambiguities. This research was supported by the Stanford Workshop on Factor Markets. I also benefited from the use of research facilities at Kyoto Institute of Economic Research and University of California at Davis.

[1] In this paper the term "labor-managed firm" (LMF) is used to designate entities that are variously called "worker-owned firms," "employee-managed firms," or "producer cooperatives."

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0002$01.50

of restructuring an ailing or bankruptcy-prone entrepreneurial firm. Second, when the LMF becomes successful, it tends to be assimilated to the entrepreneurial modus operandi. These aspects of the LMF's life cycle have been well documented empirically since the Webbs (1920) and Tugan-Baranovsky (1921). More recent records of the same phenomena are discussed by many papers collected and cited in Jones and Svejnar (1982) and Stephen (1982*a*).

My analysis follows Tugan-Baranovsky's in identifying the availability of hired wage workers as leading to the dissolution of LMFs. This effect of hired wage workers on the LMF has also been studied more recently (Ben-Ner 1982; Stephen 1982*b*). However, the scenario given by these authors is limited to the dissolution phase of a successful LMF in the long run. To explain why the LMF existed in the first place would have required them to argue that either the LMF's objectives or the member workers' preferences had changed over the course of the LMF's life cycle from the formation phase to later, more successful periods. Otherwise, since their model is static, their scenario implies only the nonexistence of LMFs. I propose here a unified framework in which both the establishment and the dissolution of the LMF can be explained as outcomes of the same LMF program with the same worker preferences but under different economic conditions.

My approach is to consider the LMF in uncertain market environments and to integrate the short- and long-run aspects of the LMF. I overcome the interpretive difficulties of previous models, which were based on LMF per capita wage maximization, by incorporating the risk-averse worker's expected utility maximization. The interactions of income risk and market uncertainty permit a clarification of the relationship between neoclassical profit maximization and economic conditions consistent with the LMF's existence. The same framework also allows an emphasis on the effect of capital-market imperfections on the economic feasibility of an LMF.

Two facets of a developed capitalist economy are represented here. One is the competitive labor market in which labor service is auctioned at a market-clearing wage rate in each state of nature. In my analysis, the existence of such a neoclassical labor market provides a key conduit through which the LMF interacts with the rest of the market economy.[2] The other facet concerns the capital market. As usual, a capitalist firm is viewed as an expected profit maximizer; this implies that the firm has access to perfect markets in lending and borrowing capital across various states of nature. As a comparative benchmark, I shall first discuss the case in which LMFs have the same access to perfect capital markets as capitalist firms have. To gain a better understanding of LMFs, however, I shall also consider the case of imperfect capital markets. Empirical observations suggest that LMFs face discriminatory imperfections in obtaining loans from private banks. (For case studies, see Jones and Svejnar [1982] and Stephen [1982*a*].) The problems of imperfect capital markets are pertinent to the existence of bona fide LMFs in the capitalist economy.

---

[2] The economic consequence of an alternative labor market regime is discussed in Sec. VII

My maintained assumptions and conventions are as follows: (i) A state of nature is denoted by s. Although it is uncertain which s occurs, the probability distribution of s is common knowledge. In interpreting the model, it is useful to think that the probability distribution of states is analogous to the textbook notion of neoclassical long-run expectations and that each state is a particular short-run realization. I shall thus use the term "ex ante" interchangeably with "long run" and "ex post" interchangeably with "short run." (ii) All workers are homogeneous and their work hours per day are institutionally fixed (8 hours per day); "wage" means "wage income" per work day. $w(s)$ denotes the market-clearing wage in state s in the competitive labor market and is the short-run opportunity cost of labor in s. It is assumed that there exists a "reservation" wage income $\underline{w}$ such that $w(s) \geqq \underline{w} > 0$ for all s. (iii) Labor input L is the only variable input in the short run, and the firm's technology requires fixed capital inputs of $\underline{K}$ units. Suppressing $\underline{K}$, we express the firm's
$R(L(s), s)$, where $L(s)$ is the number of
impose regularity conditions $R'(\infty, s)$ -
Thus the firm must have at least one
outputs.[3]

Given these assumptions, the entrep
is defined as the expected profit-maxim
profit in s is defined as

$$\Pi(s) = R(L(s), s) - w(s)L(s). \tag{1}$$

It is assumed that the PMF has access to a perfect capital market, enabling it to make optimally scheduled state-contingent payments on the rental cost of $\underline{K}$. Denote these debt payments by $D(s)$. Generally, the PMF pays a small $D(s)$, or even borrows $D(s)$, in low-revenue states, and pays a larger $D(s)$ in high-revenue states, so that on the average it pays $ED(s) = D$.[4] The PMF maximizes $E\Pi(s)$ and the bankruptcy constraint is $\max E\Pi(s) \geqq D$. Because workers can be hired from the spot market in each state,

---

[3] The number 1 is just for simplicity and stands for the minimum personnel required for production using $\underline{K}$.

[4] Throughout the paper the expectation operator E is across s.

$$\max E\Pi(s) = E[\max \Pi(s)]. \tag{2}$$

I shall now turn to the description of the LMF.

II. A Model of the LMF

The LMF is viewed here as a production coalition of member workers with a mutually binding set of contracts. Let M denote the LMF's membership size. In organizing the LMF, M members agree on policies for employment and remuneration as enforceable state-contingent contracts. Contract-based membership defines the worker's property right in the firm and identifies the LMF as being more than just the production function plus a bankruptcy constraint. Assume that membership size adjusts in the long run but is fixed in the short run. M will be determined in a manner that is consistent with the probability distribution of s.

Since the absolute size of production surplus is largely determined by output market conditions, the desired membership size will change as the long-run market conditions change. But there are several other economic factors that are also important in the LMF's membership decision. First and most important, a successful LMF has an incentive to have a small M. This arises because the per capita share of production surplus tends to be inversely related to M. This incentive to reduce M can be either reinforced or mitigated depending on the LMF's ability to earn a production surplus in the long run and also on the ease with which the LMF can substitute wage hires for members. It is in this regard that the LMF's interactions with outside labor markets become critical in inducing its own eventual demise as a labor-managed organization. We may tentatively conclude that if the long-term prospect of the LMF's earnings is good and if wage hires are readily available, then the incentive to reduce M will remain strong.

Second, if M is too small, the per capita debt burden can be severe. But the concern for the debt burden is most acute in an environment where long-term LMF earnings prospects are not good. When the PMF is faltering, workers may be able to save the enterprise by invoking the membership solidarity and by sharing the burden of the bankruptcy risk. The members' incentive to spread the per capita burden of debts (i.e., the bankruptcy risk) is a force that maintains the organization as labor managed. The decision on M therefore depends on the degree of the worker's risk aversion and on the extent to which the LMF can shift risks to outside creditors. Capital market conditions thus affect the LMF's membership size.

Third, the LMF may decide to furlough some of its members temporarily if business conditions worsen. Since layoffs create short-run income uncertainty to an individual member, the LMF will find it beneficial to create a mutual insurance to cover such risks. To insure the members fully against layoff risks, the LMF must redistribute the short-run earnings of all members in an optimal fashion between the employed and furloughed members. Such an insurance arrangement is tantamount to providing optimal layoff compensation out of the LMF's short-run wage funds. The LMF's short-run policy will therefore prescribe maximization of its wage funds in each state, according to the rule of marginal revenue productivity. To the extent that the LMF can achieve this, the LMF's earnings capacity will be enhanced and its dependency on outside capital markets will be lessened. The LMF's ability to provide optimal self-insurance in the short run is therefore another factor in determining the desired membership size.

Empirically it has been observed that the LMF's work force usually is a mixture of member workers and some wage workers. In my model the LMF expands production in the short run by hiring nonmember workers on a temporary basis when business conditions are good. To retain the character of labor management, however, it is generally thought that the LMF members must constitute a significant proportion of the long-run work force, although precise quantification is lacking in the existing LMF literature. In my analysis I shall say that the LMF degenerates into a PMF if M
a de facto entrepreneurial manager h
spot markets. Obviously, $M = 1$ is p
transformation of the LMF into a de

My formulation underscores the wo
employment stability, which has been
mary impetus for workers to organize
I make the model slightly more formal. Be
used in the paper: $v(s)$ = the wage remuneration received by an employed member of the LMF in s; $c(s)$ = the compensation paid by the LMF to a temporarily furloughed member in s; $m(s)$ = the number of members employed by the LMF in s; $n(s)$ = the number of wage workers hired temporarily by the LMF from the spot labor market at $w(s)$.

The short-run membership constraint is given by

$$1 \leqq m(s) \leqq M \tag{3}$$

for each s. Because all workers are homogeneous, the LMF lays off its members randomly with probability equal to min $\{1, 1 - [m(s)/M]\}$. Each member's mean utility in s is given by

$$U(s) = \left[\frac{m(s)}{M}\right]u(v(s)) + \left[1 - \frac{m(s)}{M}\right]u(w(s) + c(s)), \tag{4}$$

reflecting the fact that he can obtain temporary jobs elsewhere at $w(s)$. The worker's utility u is assumed to obey the usual conditions: $u' > 0$, $u'' < 0$, $u'(0) = \infty$, $u'(\infty) = 0$, and $|u(0)| < \infty$. The member's expected utility is then defined as $EU(s)$, where E is taken across s. The expected utility constraint for the LMF is

$$EU(s) \geqq Eu(w(s)). \tag{5}$$

The raison d'être of the LMF is that it can generate $EU(s)$ strictly higher than $Eu(w(s))$. Note that this requirement allows for the possibility that the member receives $v(s)$ lower than $w(s)$ in some states. Indeed, I later emphasize that the willingness of the members to accept occasionally low remunerations below their short-run opportunity income $w(s)$ is crucial to the existence and operation of the LMF. Finally, it is convenient to define for each s

$$\sigma(s) = R(m(s) + n(s), s) - w(s)n(s) - v(s)m(s) - c(s)[M - m(s)], \tag{6}$$

to be used below in specifying the LMF's bankruptcy constraint.

The LMF is then a firm that maximizes $EU(s)$ subject to (3), (5), and appropriately specified bankruptcy constraints that depend on capital market conditions. In what follows, the feasible set of LMF programs is assumed to be nondegenerate; the constraint (5) will not be explicitly mentioned every time a program is written. The mathematical solution to an LMF program may admit a constraint-binding case such that $M = 1$ or $EU(s) = Eu(w(s))$. A meaningful solution that indicates a bona fide LMF, however, is a solution such that $M > 1$ and $EU(s) > Eu(w(s))$, and I restrict attention to such solutions.[5] Having described the basic setup of the model, I go to the benchmark case in which the LMF has access to perfect capital markets and in which the LMF can distribute its production surplus optimally among the members. An optimal LMF thus described can then be modified to give a more realistic explanation of the LMF's life cycle.

---

[5] Although $EU(s) = Eu(w(s))$ fails to generate active motives for the workers to establish an LMF, it preserves the status quo incentives of the existing LMF. Provided that one is careful about the asymmetry in active incentive, it is straightforward to admit such a boundary case to the interpretation of the LMF. It is only for clarity that I require the strict inequality for a meaningful solution.

III. An Optimal LMF

It is useful to think that the LMF deals with two capital markets: internal wage funds and an external capital market. The internal wage funds function as short-run risk sharing among the members in alleviating income uncertainty due to temporary layoffs. As I shall show presently, it is instrumental in inducing short-run productive efficiency for the LMF. The external capital market, on the other hand, affects the LMF's ability to shift income risks across various short-run states. For reasons already stated, they both are important factors for the LMF's decision on M.

To illustrate the role of $c(s)$, suppose that in a given state the member has a risk of being furloughed. If employed, the member receives $v(s)$; if furloughed, he receives $c(s)$ from the LMF plus $w(s)$ from a temporary job. The member, being risk averse, will then desire to equalize the marginal utility of income between LMF employment and the temporary job. Optimal short-run risk sharing hence requires $c(s)$ such that

$$v(s) = w(s) + c(s) \tag{7}$$

if $m(s) < M$.[6] This layoff compensation scheme constitutes LMF self-insurance because it is entirely financed within the LMF's short-run wage funds by pooling all members' earnings in s.[7]

Substitute (7) into (4) and (6) to obtain

$$U(s) = u(w(s) + c(s)), \tag{8}$$

$$\sigma(s) = R(m(s) + n(s), s) - w(s$$

Equation (9) shows that the LMF's sho
parts: a quasi-fixed payment $c(s)M$ to
wage bill for the work force evaluate
$w(s)$. Since $w(s)$ is market determined,
$c(s)$ in (8). This implies maximization of $R(m$
$+ n(s)]$ in (9), which means that the marginal revenue product rule of

---

[6] If $m(s) = M$, $c(s)$ is defined as a slack variable such that $c(s) \equiv v(s) - w(s)$

[7] Suppose that a contract says $w(s) > c(s) + k$ and $m(s) < M$. An insurance policy that requires an ex ante payment from each member of $\{1 - [m(s)/M]\}[w(s) - c(s) - k]$ with full indemnification $w(s) - c(s) - k$ to each member laid off, breaks even in s. It leaves LMF's total remunerations unchanged at $m(s)w(s) + [M - m(s)]c(s)$. This scheme is preferred by the risk-averse member and is implemented by a contract employing $m(s)$, but with payment $w^*(s)$ to the employed and $w^*(s) - k$ to the furloughed, where $w^*(s) = [m(s)/M]w(s) + \{1 - [m(s)/M]\}[c(s) + k]$. A similar logic applies if $w(s)$ is less than $c(s) + k$ to begin with. In that case, the member prefers layoff to LMF employment; a premium rate of $m(s)/M$ should be charged for full coverage of $c(s) + k - w(s)$.

employment holds insofar as the LMF has access to the spot labo market. Observe

$$R'(m(s) + n(s), s) = w(s). \tag{10}$$

$$\text{Max } R(m(s) + n(s), s) - w(s)[m(s) + n(s)] = \max \Pi(s). \tag{11}$$

Efficient employment decisions by the LMF take into account the member's short-run opportunity cost of LMF employment, that is, $w(s)$. The upshot is that both the LMF and PMF have the same employment and output supply in their common production states.[8]

LMF self-insurance eliminates short-run income risk but leaves the members exposed to long-run income risk. To model this I turn to the benchmark case in which the LMF has access to the same perfect capital market as the PMF. The LMF's bankruptcy constraint is then

$$E\sigma(s) \geqq D. \tag{12}$$

Formally, the program is to

(L-1) maximize $Eu(w(s) + c(s))$ subject to (3) and (12).

The $E\sigma(s)$ constraint means that the LMF can smooth out the $v(s)$ fluctuations by shifting risks to risk-neutral creditors. I thus obtain a constant income path:[9]

$$w(s) + c(s) = \bar{v} \tag{13}$$

for all s. The LMF member is guaranteed a stable income and utility stream; $EU(s) = u(\bar{v}) = u(Ew(s) + Ec(s))$. Maximizing $EU(s)$ becomes equivalent to maximizing $Ec(s) + Ew(s)$, but $Ew(s)$ is market determined. Hence, (L-1) reduces to maximizing only $Ec(s)$. Since the bankruptcy constraint will be binding in the solution, I can further write

$$Ec(s) = \frac{E\{R(m(s) + n(s), s) - w(s)[m(s) + n(s)] - D\}}{M}. \tag{14}$$

Then, from (12), for a given membership size M,

$$\max Ec(s) = \frac{E[\max \Pi(s) - D]}{M} = \frac{\max E\Pi(s) - D}{M}. \tag{15}$$

---

[8] That is, $L(s) = m(s) + n(s)$, where $L(s) = \arg[\max \Pi(s)]$ of (1), and $m(s) + n(s) = \arg[\max \sigma(s)]$ of (6) given M. The shutdown state, however, differs between the LMF and the PMF. This is because the member's $v(s)$ does not necessarily coincide with $w(s)$.

[9] This is an extension of the previous self-insurance argument combined with the "Azariadis and Baily" theorem on fixed wage income. When members are fully employed by the LMF, $c(s)$ again is defined as a slack variable $\bar{v} - w(s)$.

From this I immediately have the following proposition.[10]

Proposition 1: In (L-1) the LMF guarantees $u(\bar{v}) > Eu(w(s))$; and

$$M = 1 \text{ and } \bar{v} > Ew(s) \text{ if and only if } \max E\Pi(s) > D; \qquad \text{(i)}$$

$$1 \leqq M < \infty \text{ and } \bar{v} = Ew(s) \text{ if and only if } \max E\Pi(s) = D; \qquad \text{(ii)}$$

$$M = \infty \text{ and } \bar{v} < Ew(s) \text{ if and only if } \max E\Pi(s) < D. \qquad \text{(iii)}$$

In an environment where the PMF can operate profitably on the average, the LMF can increase its per member remuneration if it replaces its members by hired wage workers. This is because the average wage bill paid to the hired worker is strictly less than the member's guaranteed remuneration. The reduction in the LMF membership can occur through retirements and turnovers. Departing members will not be replaced as long as $E\Pi(s)$ exceeds D. The incentive for the LMF to reduce its membership size continues in this way until only one member remains, all other workers being hired from spot markets. The result is the generalization of Tugan-Baranovsky, Stephen, and Ben-Ner.[11] This logic of LMF degeneration is different from the conventional explanation, which relies on an (implausible) assumption that the LMF's capital cost is negligible.[12]

The proposition above also says that the existence of a bona fide LMF is consistent only with an environment in which the PMF is expected to become insolvent. To put it differently, when the PMF is likely to go bankrupt, the firm may survive by becoming an LMF. This is because the worker's risk aversion to income-cum-employment risks

---

[10] Let us prove that $u(\bar{v}) > Eu(w(s))$. It
$\Pi(s) < D$, we only need to find M such

$$|Ec(s)| < |\tfrac{1}{2} \; [u''(L$$

But $Ec(s) < 0$ and $Ec(s) \to 0$ as $M \to \infty$
must hold. This establishes that the exp

[11] Since their wage-maximizing mode
defined as a firm that can pay $v > w$
successful. On the other hand, if $\Pi < D$ in their
incentive to start an LMF. Therefore, their model implies only the [illegible] tence of a bona fide LMF.

[12] In the conventional story, which does not admit the possibility of hired wage workers, the LMF maximizes per capita income $v = [R(M) - D]/M$. It determines M by $R'(M) = [R(M) - D]/M$. If R is concave, M is an increasing function of D, and $M \to 1$ as $D \to 0$. A natural adaptation of the conventional LMF model to our uncertainty framework is to let the LMF solve (L-1) under the condition that it does not hire wage labor, i.e., $n(s) = 0$ for all s. Then M will be determined implicitly by $ER'(m(s), s) = \bar{v}$ $E\{R(m(s), s) + w(s)[M - m(s)] - D\}/M$. Once again, $\partial M/\partial D > 0$ and $M \to 1$ as $D \to 0$ provided that R is concave in m for each s. Without hired wage workers, the conventional LMF model predicts $M = 1$ if and only if D can be made negligible. In contrast, our proposition shows that whenever $E\Pi(s) > D$, $M = 1$ even if D is significantly large. Clearly, it is the interactions with the outside wage-labor market that degenerate the LMF in (L-1).

enables the LMF to attain the wage bill reduction that is necessary for the firm's survival. Further, each member of the LMF finds it beneficial to share the firm's debt burden with as many fellow workers as possible because the firm's expected "shadow profits" are negative.

The (L-1) program suggests that the long-run existence of the LMF is inversely related to the expected shadow profitability of the firm. The workers of a failing PMF can collectively have an incentive to buy it out and reorganize it into an LMF. If the long-run shadow profitability of the LMF improves, however, members lose the incentive to maintain membership size. In a long-run steady state with positive shadow profits, the LMF will dissolve into a de facto PMF. To emphasize, the LMF in the present context is productively as efficient as the PMF throughout its life cycle. The existence of the LMF merely underscores the PMF's inability to avoid bankruptcy through wage cuts and other labor concessions; the LMF dissolves into the PMF in spite of, and because of, its very success in the long run.

An interpretive anomaly in (L-1) is that once the LMF becomes feasible, its desired membership size becomes unbounded. This unboundedness in M results from the insurance operation of the LMF, not from the actual production of output. The actual work force employed, $m(s)$, is always finite under the usual assumption of a concave revenue function. Because optimal risk sharing–cum-shifting makes the member completely indifferent between LMF employment and jobs elsewhere in any state, however, the LMF as an insurance company effectively possesses a constant-returns-to-scale technology in M. Thus, if $Ec(s) < 0$, which is necessary for the LMF's existence, $E\sigma(s)$ increases without limit as M becomes unboundedly large. All of this implies that the LMF's membership size will be delimited if it is not possible to insure fully the member's income either in the short run (within each s) or in the long run (across s). As I shall show in the next two sections, however, incomplete self-insurance and imperfect capital markets have their own effects on the operations of the LMF as they respectively serve to limit the potential membership size. Such extensions of the basic model to imperfect environments are not only more realistic but also helpful to understanding the government's practice of underwriting capital loans when the failing PMF is bought out by its employees.

IV. Incomplete Self-Insurance

Proposition 1 states that a necessary condition for LMF existence is $Ec(s) < 0$. It means that members accept occasional $v(s) < w(s)$ to obtain a long-term gain $u(\bar{v}) > Eu(w(s))$ from the LMF. A negative $c(s)$ also means that furloughed members channel portions of their outside earnings back to the LMF to ensure the optimum self-insurance. Obviously, enforceability problems will arise in such transfer arrangements unless the LMF is founded on a strong feeling of member solidarity. Practical difficulties nevertheless exist in correctly assessing $c(s)$ if there is uncertainty as to the jobs available elsewhere; if taxes and benefits of government-financed dole are not completely experience rated; if $w(s)$ includes the value of home production and leisure; if workers have sources of unearned income, etc. Generally, ex post transfers are likely to be less efficiently assessed, the more heterogeneous workers and firms become. It is therefore realistic to consider the case of LMF in which such transfers are suboptimal. I analyze here a simple case where $c(s) = 0$.[13] To isolate the effects of incomplete self-insurance, however, I continue to assume that the LMF has access to risk shifting across states.

Formally, the LMF's program is modified to

(L-2) maximize $EU(s)$ subject to (3) and (12)

where $U(s)$ and $\sigma(s)$ are given by

$$U(s) = \left[\frac{m(s)}{M}\right]u(v(s)) + \left[1 - \frac{m(s)}{M}\right]U(w(s)) \tag{16}$$

and

$$\sigma(s) = R(m(s) + n(s), s) - w(s)n(s) - v(s)m(s). \tag{17}$$

Since the across-state risk shifting is still
of smoothing the member's remun

$$\bar{v} =$$

constant for all s. Clearly, $\bar{v}$ can be
fluctuates.

Since risk-neutral creditors abso
fluctuations, the LMF's long-run pe
marized by using (11). After some manipu
Appendix, it is possible to confirm

$$EU(s) \leqq u\left(Ew(s) + \left(\frac{1}{M}\right)E[\max \Pi(s) - D]\right). \tag{19}$$

from which several observations follow. In the environment in which the PMF is profitable, no bona fide LMF can exist because the right-hand side of equation (19) will then be maximized at $1 = m(s) = M$. An LMF will be organized only if the PMF is bound to be insolvent. In

---

[13] It is technically straightforward to extend the analysis to a bounding case such as $\underline{c} \leqq c(s) \leqq \bar{c}$

(L-2), however, if the PMF is gravely insolvent it does not pay to reorganize it as an LMF, for the requirement $EU(s) > Eu(w(s))$ cannot be met if the expected loss $D - E[\max \Pi(s)]$ is too large. This is in contrast to (L-1) in which the optimal self-insurance coupled with perfect capital markets always ensured economic viability of the LMF in the event of PMF insolvency. Also, in (L-2) M will not be unbounded; because $m(s)$ is finite, an infinitely large M reduces the probability of LMF employment, $m(s)/M$, to zero, and hence the member obtains just $Eu(w(s))$. In other words, the lack of optimal layoff compensations makes the income differential due to furloughs riskier. The risk-averse members will try to mitigate this income risk by raising the LMF employment probability, which implies that M cannot be too large.

The foregoing intuitive explanations are made precise in the Appendix; we summarize the results below:

Proposition 2: The solution to (L-2) entails:

$$\bar{v} > Ew(s) \text{ and } M = 1 \text{ if and only if } \max E\Pi(s) > D,$$

$$\bar{v} < Ew(s) \text{ and } 1 < M < \infty \text{ if } \max E\Pi(s) < D,$$

$$\bar{v} = Ew(s) \text{ and } 1 \leqq M < \infty \text{ if } \max E\Pi(s) = D.$$

The incompleteness of self-insurance eliminates the constant-returns-to-scale aspect of the insurance operations, and the size of an LMF becomes delimited by market conditions and the production technology. Despite the absence of optimal transfers between the employed and furloughed members, a bona fide LMF can still be viable in (L-2) because members are willing to accept occasional subpar remunerations in return for long-term income stability.

The lack of efficient risk sharing, however, induces certain inefficiencies in the LMF's ex post production. Because the "price" adjustment in terms of wages and layoff compensation is now limited, the members' risk aversion will cause the short-run employment level $m(s)$ to become larger than the productively efficient level. These short-run "quantity" adjustments in $m(s)$ can be seen as follows. First, when $\bar{v} < w(s)$, ex ante efficient risk sharing would have required the furloughed members to contribute their outside earnings to the LMF. In the absence of such ex post sendbacks, more members are obliged to engage in LMF production for $\bar{v}$ below $w(s)$. On the other hand, when $\bar{v} > w(s)$, the member incurs an income loss from being furloughed. In the absence of optimal layoff compensations to mitigate this risk, the LMF increases the employment probability by retaining members beyond the efficient separation level. Either way, LMF policy entails short-run overemployment. Efficient short-run employment is fortuitous only in states where $\bar{v} = w(s)$.[14]

Because of this short-run overemployment, the LMF in (L-2) will not exactly duplicate the employment-cum-output path of the PMF counterpart. But, unlike Ward's (1958), Vanek's (1970), and Meade's (1972) Illyrian firm, which reduces output in response to a rise in its output price, my LMF always exhibits a normal supply response. To see this, consider a competitive case such that $R(\cdot, s) = sf(\cdot)$, where s is the parametric output price and f the firm's production function. It can then be demonstrated that output supply, $m(s)$, and $n(s)$ are all increasing functions of s. This is because across-state risk shifting enables the LMF to maximize implicitly $E\sigma(s)$ for a given level of $EU(s)$. This can be achieved by planning to sell more output in high-s states.

V. Imperfect Capital Markets

Others have argued that financial markets are biased against LMFs in favor of PMFs. To incorporate capital market imperfections that the LMF might face, I specify in this section a state-by-state bankruptcy constraint as[15]

$$\sigma(s) \geqq D \tag{20}$$

for each s, where $\sigma(s)$ is given by (9
that the LMF can implement op
short-run efficiency conditions suc
from the neoclassical norm will be d
fections.

The new program is to

(L-3) maximize $Eu(w(s) + c(s))$ subje

Because (20) is a binding constraint in each state,

$$\max Eu(w(s) + c(s)) = E \max u(w(s) + c(s)) = Eu(w(s) + \max c(s)) \tag{21}$$

---

[14] All of this can be made precise by inspecting the first-order conditions of (L-2). They are: (i) $n(s) > 0$ only if $m(s) = M$, (ii) $R'(m(s), s) < w(s)$ iff $m(s) < M$; (iii) $E[R'(m(s) + n(s), s) - \bar{v}]m(s) = 0$.

[15] A tacit assumption has been that the individual worker is unable to obtain a long-term state-contingent consumption loan unless the employer firm arranges it for him. Hence, the firm's difficulty in borrowing across states directly translates into the worker's inability to make interstate substitutions in his consumption-income stream

Furthermore, $c(s) = \{R(m(s) + n(s), s) - w(s)[m(s) + n(s)] - D\}/M$, and

$$\max c(s) = \frac{\max \Pi(s) - D}{M}. \tag{22}$$

In the rest of this section I economize on notation by writing $\Pi^*(s) \equiv \max [\Pi(s) - D]$. Therefore (L-3) becomes a program to choose M so as to

$$\max_{M} Eu\left(w(s) + \frac{\Pi^*(s)}{M}\right). \tag{23}$$

From (23) it is evident that any meaningful LMF solution in (L-3) will have a finite M, so that $Eu(\cdot) > Eu(w(s))$ can be met. Also, if $\Pi^*(s) > 0$ in every s, then $M = 1$ and de facto entrepreneurial production prevails. Unlike (L-1) and (L-2), however, the LMF will never be organized if $\Pi^*(s) < 0$ for all s. Because the LMF here lacks access to interstate debt financing, mutual risk sharing becomes a more central motive for the establishment of an LMF. I elaborate here on the role of risk aversion and income fluctuations in the determination of M.

The LMF can reduce the member's income risk due to the across-state variability of $\Pi^*(s)$ by enlarging M. But a larger M means a lower per capita income in each state. Further, because both $w(s)$ and $\Pi^*(s)$ fluctuate across states, the member's expected utility crucially depends on their covariance. The extent to which income smoothing takes place via M depends on how $E[\Pi^*(s) + w(s)]$ and cov $[w(s), \Pi^*(s)]$ get translated into the mean and variance of the member's utility. The need for smoothing will therefore be greater as the member becomes more risk averse. One can see these effects on M more clearly if one checks the derivative of $Eu(w(s) + [\Pi^*(s)/M])$ with respect to M. Then,

$$\operatorname{sgn}\left\{\frac{\partial Eu(w(s) + [\Pi^*(s)/M])}{\partial M}\right\} = -\operatorname{sgn}\left\{Eu'\left(w(s) + \left[\frac{\Pi^*(s)}{M}\right]\right)\Pi^*(s)\right\}. \tag{24}$$

The sign term of the right-hand side can further be expanded as

$$Eu'\left(w(s) + \frac{\Pi^*(s)}{M}\right) \cdot E\Pi^*(s) + \operatorname{cov}\left[\Pi^*(s), u'\left(w(s) + \frac{\Pi^*(s)}{M}\right)\right]. \tag{25}$$

It is not possible to determine a priori the sign of this covariance term.[16] If $\partial Eu(\cdot)/\partial M > 0$ as $M \to \infty$, M becomes unbounded. That is,

$$\begin{aligned} 0 &\gtreqless Eu'(w(s))\Pi^*(s) \\ &= Eu'(w(s)) \cdot E\Pi^*(s) + \operatorname{cov}[u'(w(s), \Pi^*(s))]. \end{aligned} \tag{26}$$

---

[16] If u' is linear, $\operatorname{cov}\{\Pi^*(s), u'(w(s)) + [\Pi^*(s)/M]\} = [\operatorname{var} \Pi^*(s)]/M + \operatorname{cov}[\Pi^*(s), w(s)]$. Even then, the sign of cov $[\Pi^*(s), w(s)]$ cannot be determined a priori.

On the other hand, $M = 1$ occurs if and only if $\partial Eu(\cdot)/\partial M < 0$ at $M = 1$:

$$\begin{aligned} 0 &\leqq Eu'(w(s) + \Pi^*(s))\Pi^*(s) \\ &= Eu'(w(s) + \Pi^*(s))E\Pi^*(s) + \operatorname{cov}[u'(w(s) + \Pi^*(s)), \Pi^*(s)]. \end{aligned} \tag{27}$$

Therefore, a bona fide LMF as an interior solution ($1 < M < \infty$) exists if and only if neither (26) nor (27) holds.[17]

Summing up, we obtain:

Proposition 3: The LMF in (L-3) has a finite M that is uniquely determined by equating (26) to zero. When the LMF exists, $\Pi^*(s)$ fluctuates around zero and satisfies the following inequality conditions:

$$\begin{aligned} &\frac{\operatorname{cov}[u'(w(s) + \Pi^*(s)), \Pi^*(s)]}{Eu'(w(s) + \Pi^*(s))} \\ &\quad < -(E\Pi^*(s)) \\ &\quad < \frac{\operatorname{cov}[u'(w(s)), \Pi^*(s)]}{Eu'(w(s))}. \end{aligned}$$

Clearly, depending on the sign and magnitude of the covariance terms, the existence of the LMF is consistent with either a positive or a negative $E\Pi^*(s)$. It also turns out that the interpretations of the LMF in (L-3) differ as cov $[w(s), \Pi^*(s)]$ be[illegible] discuss the two separate cases b[illegible]

Case 1: cov $[w(s), \Pi(s)] > 0$

If the firm's business perform[illegible] roeconomic fluctuations in the [illegible] and cov $[w(s) + \Pi^*(s), \Pi^*(s)]$ [illegible] covariance conditions imply in p[illegible] is consistent only with $E\Pi^*(s)$ [illegible] previous findings but can be explained [illegible] that $\Pi^*(s)$ fluctuates between positive and nega[illegible] < 0, the member's expected income is less than $Ew(s)$. Additionally, because members lack recourse to outside risk shifting, the positive covariance between $w(s)$ and $\Pi^*(s)$ only increases the member's income risk. Consequently, the risk-averse worker will not be motivated to set up an LMF. The situation is opposite with $E\Pi^*(s) > 0$, in which case workers have incentives to participate in the LMF because they expect to obtain income higher than $Ew(s)$. Further, if $\Pi^*(s)$ fluctuations are sufficiently volatile, risk-averse workers will find it beneficial to reduce the per capita $\Pi^*(s)$ variability by a factor of M.

---

[17] In (23) the maximand $Eu(\ )$ is not necessarily concave in M. Nevertheless, $Eu(\ )$ is locally concave in the neighborhood of M such that $\partial Eu(\cdot)/\partial M = 0$. Hence, by inspecting the respective asymptotic values of $\partial Eu(\cdot)/\partial M$ as $M \to \infty$, we can verify if there exists an interior solution ($1 < M < \infty$) to (23). Also, (26) and (27) are mutually consistent in the sense that $Eu'(w(s))\Pi^*(s) \leqslant 0$ implies that $Eu'(w(s) + \Pi^*(s))\Pi^*(s) < 0$.

[18] *This is to say that the wages vary procyclically with the performance of the economy.* It has also been known that the profit share of GNP has been much more volatile than the wage share. Note also that cov $\{\Pi^*(s), w(s) + \Pi^*(s)\} > 0$ does not imply cov $\{\Pi^*(s), w(s) + [\Pi^*(s)/M]\} > 0$ for $M > 1$.

When a positive $E\Pi^*(s)$ is a necessary condition for the LMF's existence, workers will be less concerned with income losses due to the firm's bankruptcy. Their more immediate motive is to obtain a larger share of the firm's quasi rents and to increase $Eu(\cdot)$ above $Eu(w(s))$. The practical manifestation of this type of LMF is perhaps found in partnerships and professional group practices—for example, accountants, lawyers, medical practitioners, and financial intermediaries.

Case 2: $\text{cov}\,[w(s), \Pi^*(s)] < 0$.

Since the short-run $w(s)$ will be determined largely by the overall condition of the labor market, microeconomic disturbances particular to a given firm can cause a negative $\text{cov}\,[w(s), \Pi^*(s)]$. Such a scenario includes industries that are declining more rapidly than the cross-section average of the economy. More generally, a negative $\text{cov}\,[w(s), \Pi^*(s)]$ will be observed if the firm's output market—that is, $R(\cdot, s)$ function—is statistically more stable than the labor market, that is, $w(s)$. Now suppose $\text{cov}\,[w(s), \Pi^*(s)] < 0$ and $\text{cov}\,[w(s), \Pi^*(s) + w(s)] > 0$.[19] Then there is a neighborhood around $E\Pi^*(s) = 0$ in which the inequalities of proposition 3 can be met. The LMF's existence is again consistent with $E\Pi^*(s) > 0$; both risk reduction and rent seeking can account for such an LMF. If the firm's debt burden is not excessive, workers can be motivated to restructure the insolvent PMF.

VI. Qualifications

The detailed features of my analysis obviously depend on the simplifying assumptions, but the principles that it illustrates are general. In this section I suggest some interpretive qualifications and extensions that follow directly from the foregoing analysis.

A. *Imperfect Capital Markets*

When there exist simultaneous imperfections in the capital market and in the implementation of self-insurance, the technical analysis becomes less tractable. Clearly short-run overemployment occurs, and the potential size of the LMF will be delimited. Surprisingly, however, simultaneous imperfections in the capital markets do not alter the essential implications of the expected utility maximization for the LMF's short-run behavior. Such an LMF might deviate noticeably from the neoclassical norm only in the very low-revenue states in which the LMF would still operate but its twin PMF would shut down.[20]

---

[19] Because $\Pi^*(s)$ is the total profit of the firm as a whole and $w^*(s)$ is the per capita wage income, $\text{cov}\,[\Pi^*(s), w(s) + \Pi^*(s)] > 0$ is not inconsistent with $\text{cov}\,[w(s), \Pi^*(s)] < 0$

B. *Internal Financing*

Vanek (1975) argues that LMFs in a capitalist economy are likely to rely more heavily on internally financed investments than PMFs. He explains this tendency as resulting from discriminations in capital markets, which are generally mistrustful of labor-managed organizations. Internal financing reduces external debts D in the model. To the extent that it eases the LMF's bankruptcy constraint, internal financing reinforces the incentive to reduce M.

C. *Member/Wage Labor Substitutability*

The dissolution episode of a successful LMF has been made dramatic in our scenario because of the assumed worker homogeneity. It is more plausible that the long-term members and temporary wage hires differ in their labor quality, especially if the LMF's rents are generated by its organization-specific human capital. The LMF's revenue function should then be $R(m(s), n(s)$ [illegible] Because the imperfect substitutability between members [illegible] able they are and the more prom[illegible] are, the larger the long-run mem[illegible] LMF.[21]

---

[20] This statement can be illustrated most [illegible] The LMF has an upward-sloping output supply [illegible] the PMF also produces. But there is a range of very [illegible] *produces and the PMF shuts down in the short run. It is in this low-output price range* that a portion of the LMF's short-run supply curve might have a negatively sloped section. This is not, however, the same kind of well-known "perversity" as is seen in the Vanek-Ward Illyrian firm. There a rise of the output price per se causes a negatively sloped supply curve everywhere. In the case being discussed, the excessive short-run debt burden obliges the LMF to employ its members at remunerations below their shadow wage. Because of the simultaneous imperfections in the internal and external capital markets, such a wage-employment rule is nevertheless consistent with $EU(s)$ maximization. When the output price rises in that range, the LMF may be able to release some of its members from the "involuntary" overemployment situation, thereby possibly reducing its output supply. See Miyazaki and Neary (1983) for details.

[21] If R is linear homogeneous in (m, n, other variable inputs), then $M = 1$ once again even with imperfect substitutability between m and n.

CPMF (contractual) to distinguish it from PMF; for the CPMF I call M the labor pool.[26]

From the structure of programs, it is clear that there exists a short-run qualitative equivalence between the CPMF and the LMF. Once M is determined, the short-run employment rule of the CPMF is essentially the same as that of the LMF under perfect capital market conditions. The comparison of the two programs suggests that if the CPMF goes bankrupt, then its dual LMF will also become insolvent. Hence, at least in theory, insofar as entrepreneurial firms offer employment contracts, it is difficult to see why pure LMFs should exist as alternatives to failing CPMFs. However, even in the regime where labor market transactions are pervasively contractual, there are reasons why the firm's wage-employment policy is likely to possess a hybrid character of the LMF and PMF.

The aforementioned qualitative duality between the CPMF and LMF is only for short-run wage-employment variations; their responses to changes in long-run parameters such as the probability distribution of s differ. In fact it can be demonstrated that under given long-run conditions, the CPMF tends to have a larger M, lower $\bar{v}$, and higher short-run layoff probabilities than the corresponding LMF. This means that a possible conflict arises between labor and management regarding the long-run aspect of wage-employment policies. Such a conflict can realistically lead to "internal bargaining" especially if the worker's long-term contractual attachment to the firm creates a firm-specific "organizational capital" to which rents accrue. In other words, the conflict is over the division of Marshallian quasi rents, and the relative bargaining power between labor and management will decide the specification of M and $\bar{v}$ in the negotiated contract.[27]

Via the bargaining process workers indirectly participate in the firm's management. But the process and the objective of negotiation much depend on the enforceability of the agreement. Enforceability problems arise whenever there exists ex post informational asymmetry between the bargaining parties. For example, management is in the position to measure more accurately workers' marginal revenue product to the firm whereas workers know better about the opportunity cost of their labor. Since monitoring the other party's conduct is usually costly and imperfect, it becomes difficult to write a contract that ensures bargaining efficiency ex ante at the same time as it prevents moral hazard ex post.

The more severe the enforceability problem is, the further removed the negotiated contract will be from the efficient frontier. A faltering long-term performance of the firm can exacerbate these enforceability issues. When labor's concession is genuinely called for, management may lack the credibility necessary for honest bargaining If that is the case, one way to facilitate a more cooperative bargaining process is to admit more direct labor participation in the firm's management. Such an organizational remedy will be particularly effective if the cause of organizational inefficiency is due to informational impactedness between the two parties.[28] In an extreme circumstance, the conversion of the firm into employee ownership may be a less costly procedure than risking bankruptcy from a protracted bargaining impasse.

---

[26] Note that $c(s)$ is absent from the original Azariadis-Baily-Holmstrom formulations; their CPMF is similar to the dual program of my (L-2.)

[27] See Aoki (1980), Svejnar (1982), and Miyazaki (1984) for models of an internal bargaining firm.

The upshot of the paradigm is then twofold. If the capitalist labor market operates as spot transactions, then the LMF may effectively reorganize the insolvent PMF. If the entrepreneurial firm recognizes the advantage of employment contracts, internal bargaining between labor and management becomes an important factor in determining the participatory content of the firm. Furthermore, if contractual enforceability becomes a serious issue, a more direct participation by labor in the firm may allay the second-best aspects of the contractual relationship. Some form of labor participation seems inevitable when the firm's performance is affected by the bilateral mode of internal bargaining.

VIII. Conclusion

I have proposed a simple but fo
cycle aspect of LMFs in the develop
also has implications for the labor
and other similar subeconomies suc
dragon. My analysis suggests that a
efficiency may induce a disintegration
economies as social institutions. Short-run productive efficiency requires the smooth reshuffling of mobile labor by means of shadow wage payments. Over the long run, successful LMFs would then exhibit propensities to become more entrepreneurial with smaller membership size, and only the less successful LMFs tend to remain as large LMFs.

The model can be extended to allow heterogeneous workers and capital-labor substitutions in the firm's production function without affecting the essential aspects of interpretation. It is important, however, to recognize that throughout the paper we have discussed the firm's performance as being completely determined by exogenous market conditions and a neoclassical production function. I have not considered the possibility that a change in the organizational mode from the PMF to the LMF itself would affect the firm's productivity. Labor relations literature indicates that workers' motivation depends partly on the workplace environment and the firm's internal structures. Many business experts consider the workers' corporate consciousness and organizational awareness significant factors for the successful implementations of QC-circles and productivity gains. I have not grappled with these internal factors. Instead I have focused on the organizational consequences of interactions between the technologically efficient firm and the outside market environment.

---

[28] Casual empiricism in current news reports certainly seems to confirm the existence of such an enforceability/credibility gap in bargaining between labor and management Besides the well-publicized cases of Chrysler and International Harvester negotiations, the airline and steel industries have faced similar problems In all these reported cases, the bargaining process is exacerbated because workers are wary of management's commitment to reinvest labor's concessions directly into their own firms and industries

Appendix

To confirm proposition 2, I define a "slack variable" $k(s)$ such that $k(s) = \bar{v} - w(s)$. Then $\sigma(s)$ can be expressed as $\sigma(s) = R(m(s) + n(s), s) - w(s)[m(s) + n(s)] - k(s)m(s)$. Since $\max \{R(m(s) + n(s), s) - w(s)[m(s) + n(s)]\} = \max \Pi(s)$, and because the bankruptcy constraint binds in the solution, we have

$$Em(s)k(s) = E\{R(m(s) + n(s), s) - w(s)[m(s) + n(s)]\} - D. \qquad \text{(A1)}$$

Because u is strictly concave, $EU(s)$ can be transformed as follows:

$$\begin{aligned} EU(s) &= E\left\{\left[\frac{m(s)}{M}\right]u(w(s) + k(s)) + \left[1 - \frac{m(s)}{M}\right]u(w(s))\right\} \\ &< Eu\left(\left[\frac{m(s)}{M}\right][w(s) + k(s)] + \left[1 - \frac{m(s)}{M}\right]w(s)\right) \\ &= Eu\left(\left[\frac{m(s)k(s)}{M}\right] + w(s)\right) \\ &< u\left(E\left[\frac{m(s)k(s)}{M}\right] + Ew(s)\right). \end{aligned} \qquad \text{(A2)}$$

By combining (A1) and (A2) we obtain

$$EU(s) < u\left(Ew(s) + \left(\frac{1}{M}\right)E[\max \Pi(s) - D]\right). \qquad \text{(A3)}$$

Since $Ew(s)$ and $\max \Pi(s)$ are market determined, the right-hand side of (A3) is maximized by $M = 1$ if $E \max \Pi(s) > D$, and $M = m(s) = 1$ is feasible in (L-2). Hence, a meaningful LMF exists only if $\max E\Pi(s) \leqslant D$.

It is also clear that the right-hand side of (A3) increases as $M \to \infty$ if $E\Pi(s) < D$ In that case, however, the right-hand side of (A3) is just one upper bound for the $EU(s)$; it needs to be checked whether $M = \infty$ constitutes a meaningful solution. It can be verified that $m(s)$ is finite for all s when $E\sigma(s) \geqslant D$. This follows from the combination of $\bar{v} > 0$, $R'(\infty, s) = 0$, and $w(s) \geqslant \underline{w} > 0$ for all s.

Now, $0 \leqq m(s) < \infty$ and $M = \infty$ mean $EU(s) = Eu(w(s))$. That is, no meaningful LMF can exist with $M = \infty$ in (L-2). Lastly, M is given by $E[m(s)R'(m(s) + n(s), s)] = Em(s)\bar{v}$, where $R'(M + n(s), s) = w(s)$ for s such that $n(s) > 0$. From this and equation (A1) it can be verified that $M = 1$ is consistent with $E\Pi(s) = D$.

References

Aoki, Masahiko. "A Model of the Firm as a Stockholder-Employee Cooperative Game." *A.E.R.* 70 (September 1980): 600–610.

Azariadis, Costas. "Implicit Contracts and Underemployment Equilibria." *J.P.E.* 83 (December 1975): 1183–1202.

Baily, Martin N. "Wages and Employment under Uncertain Demand." *Rev Econ. Studies* 41 (January 1974): 37–50.

Ben-Ner, Avner "On Cooperatives and Hired Labor: Two Step Optimization and Instability." Working Paper no 194. Davis: Univ California, Davis, Dept. Econ., July 1982.

Berman, Katrina V. *Worker-owned Plywood Companies: An Economic Analysis* Pullman: Washington State Univ. Press, 1967.

Holmstrom, Bengt. "Equilibrium Long-Term Labor Contracts." *Q J E.* 98 (suppl., 1983): 23–54.

Jones, Derek C., and Svejnar, Jan, eds *Participatory and Self-managed Firms Evaluating Economic Performance.* Lexington, Mass . Heath, 1982.

Meade, James E. "The Theory of Labour-managed Firms and Profit Sharing" *Econ J* 82 (suppl.; March 1972) 402–28.

Miyazaki, Hajime. "Internal Bargaining, Labor Contracts, and a Marshallian Theory of the Firm." *A E R* 74 (June 1984): 381–93

Miyazaki, Hajime, and Neary, Hugh M. "The Illyrian Firm Revisited" *Bell J Econ* 14 (Spring 1983): 259–70

Schlicht, Ekkehart, and von Weizsacker, Carl C
Managed Economies: The Commitm
Staatswissenschaft 31 (suppl.; 1975)

Stephen, Frank H., ed. *Performance of L*
lan, 1982. (*a*)

———. "The Economic Theory of the
mance of Labour-managed Firms Lond

Svejnar, Jan. "On the Theory of a P
(August 1982): 313–30.

Thomas, H. "The Performance of the Mondr
Participatory and Self-managed Firms Evaluating Economic Performance, edited by Derek C. Jones and Jan Svejnar. Lexington, Mass : Heath, 1982

Tugan-Baranovsky, Mikhail I. *Sotsialnyia osnovy kooperatsii* Berlin. Slowo Verlagsgesellschaft, 1921.

Vanek, Jaroslav. *The General Theory of Labor-managed Market Economies* Ithaca, N.Y., and London: Cornell Univ. Press, 1970.

———. "The Basic Theory of Financing of Participatory Firms" In *Self-Management: Economic Liberation of Man*, edited by Jaroslav Vanek. Harmondsworth: Penguin, 1975.

Ward, Benjamin. "The Firm in Illyria: Market Syndicalism." *A E R* 48 (September 1958): 566–89.

Webb, Sidney, and Webb, Beatrice. *A Constitution for the Socialist Commonwealth of Great Britain.* London: Longmans, Green, 1920.

Regulation and Industrial Organization

Andrew F. Daughety

University of Iowa

This paper examines the effects on firm behavior and industry structure of industry-wide price regulation by the state. Such regulation, wherein regulators set prices with regard to an industry-wide aggregate performance measure, has been (and still is) widely used and is shown to induce inefficiencies in firm operation. In general such regulation can result in setting prices high enough to maintain inefficient firms, and it encourages inefficient firm operation due to incentives to inflate costs and Averch-Johnson effects.

Introduction

This paper examines the effects on firm behavior and industry structure of industry-wide price regulation by the state. Such regulation, wherein prices are set with regard to an industry-wide aggregate performance measure, has been (and still is) widely used and will be shown to induce inefficiencies in firm and industry operation. Thus, this paper will address two issues associated with such regulation. First, what are the effects on individual firms of industry-wide regulation? Second, does industry-wide regulation induce or favor certain industry structures, and what are the interfirm effects of such regulation?

The salient feature of industry-wide regulation is that industries typically are composed of a number of firms of different sizes yet are subject as a group to regulatory control. A common denominator of

This work was partially supported by NSF grant SES-8218684 and by a grant from the Transportation Center, Northwestern University. Special thanks to Robert Baesemann, David Baron, Ronald Braeutigam, Benjamin Eden, Robert Forsythe, John Kennan, John Kwoka, Robert Ludke, Leon Moses, James Rosse, F M Scherer, Robert Taggert, and S. Y Wu for valuable suggestions on this and previous versions

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago All rights reserved 0022-3808/84/9205-0012$01.50

the vast majority of previous analyses of price regulation is that the regulatory restrictions in question can be examined at the level of an individual firm without reference to other firms in the industry. This reflects a view that either the analysis can be generalized to a multifirm industry via replication (i.e., all firms act the same) or the industry is comprised of one firm (the monopoly case). Examples of the former situation are the standard analyses of price supports in agriculture and minimum wage legislation in perfectly competitive labor markets (see, e.g., Hirshleifer 1976). The latter view has been examined thoroughly in the extensive Averch-Johnson (A-J) literature (see, e.g., Baumol and Klevorick 1970), wherein a firm is price regulated via a rate-of-return restriction. In both situations the analysis incorporates the notion that the firm's profit-maximizing strategies are only restricted by the regulation and that the regulation falls inescapably on the firm in question.

In general, however, a number of regulated industries are not properly modeled in the fashion described. These industries are composed of a number of firms of different sizes yet are subject as a group to regulatory control. Industry-wide regulation typically has evolved in the past when the decision to regulate a large industry recognized the cost of case-by-case review and opted instead to have the industry as a whole meet the regulatory restriction in question. For example, this was one of the main motivations of the decision by the Federal Power Commission (now the Federal Energy Regulatory Commission) to establish regional wellhead prices for natural gas production (see MacAvoy 1979). Similarly, World W
often set with reference to average
(see Harris 1945). Two modern
purchases and medicare reimburse
dairy industry production costs are
chase price for cheese products.
hospitals for medicare will be base
costs (with some adjustment for regiona
rural location, and teaching/nonteaching status of the hospital

Another perceived advantage of regulation based on an industry norm is the pooling of information presumably provided by such an exercise. For example, insurance rates are often set by state regulators with regard to average loss, expense, and profit experience by the insurance industry (see Joskow 1973; Ippolito 1979). The Interstate Commerce Commission has regulated motor carrier rates using the ratio of aggregate industry costs to aggregate industry revenue (called the operating ratio) for many years. Apparently the motivation for using this approach was a combination of desire to avoid the administrative cost of dealing individually with a large number of trucking firms, and the difficulty of setting a "reasonable" rate in a potentially volatile industry through other means (see Locklin 1972; Dobesh 1973). Other examples of industry-wide regulation include the Civil Aeronautics Board, which has used industry-averaged rate of return for airlines to judge the reasonableness of fare adjustments (see Douglas and Miller 1974) and the control of entry into the taxicab market in Chicago (see Kitch, Isaacson, and Kasper 1971).

The paper proceeds as follows. First I formalize the nature of the regulatory constraint and the regulatory scenario (the sequence of actions taken by the regulator and industry). Two constraints will be considered in detail: average cost and operating ratio. Next, I consider their application to a one-firm industry (wherein the two constraints coincide). Two results from the literature are reviewed. First, to the degree that such constraints are asymmetric in the treatment of factors of production, A-J biases will occur (see Cherry 1978; Moore 1978). Second, such constraints provide an incentive for the firm to engage in inefficient production. Thus "goldplating" à la Alchian and Kessel (1962) is not only utility maximizing, it is profit maximizing.

Next I consider the multifirm case, where I specifically allow for nonsymmetric firms. Firms will be viewed as choosing output levels so as to maximize profits, subject to the regulatory constraint. In order to accomplish this, the industry will be examined by means of a Stackelberg model with the leader facing the regulatory constraint, which, in conjunction with the reaction functions of the followers, determines industry output. In general (and depending on the specifics of the industry-wide regulatory mechanism), interfirm distortions and potentially very significant inefficiencies in the structure and performance of the industry will be shown to result from such regulation.

Industry-wide Regulatory Constraints

The previous industry examples suggest that two types of constraints have been used to regulate prices on an industry-wide basis using an average measure. The first, the average cost (AC) constraint, is formed by taking the average of the individual average costs of production. This number is then used to restrict the allowed price (via a proportionality constant). The second, the operating ratio (OR) constraint, is formed by taking the ratio of operating costs to operating revenues and requiring this average to meet or exceed some prespecified fraction.

More formally, let $\mathbf{x} \in R^n_+$ be the output vector for the n firms that comprise the regulated group;[1] firm i produces output x_i. Each firm has a cost function $c_i(x_i)$ that is assumed to be monotonically nondecreasing and differentiable (other assumptions will be made as necessary). Finally, the industry faces an inverse demand function $p(X)$ that specifies the maximum price consumers will pay for the industry's homogeneous product[2] as a function of aggregate output $X = \Sigma_{i=1}^{n} x_i$. Let α be a positive fraction ($\alpha \in [0, 1]$) set by a regulator. The constraints can be written as[3]

---

[1] $R^k_+ = \{\mathbf{x} \in R^k | x_i \geq 0, i = 1, \ldots, k\}$.

$$\text{AC: } \frac{1}{n}\sum_{i=1}^{n} \frac{c_i(x_i)}{x_i} \geq \alpha p(X),$$

$$\text{OR: } \frac{\sum_{i=1}^{n} c_i(x_i)}{\sum_{i=1}^{n} p(X) x_i} \geq \alpha.$$

For convenience, we reexpress the constraints as follows: $F(\mathbf{x}, \alpha) = h(\mathbf{x}) - \alpha p(X) \geq 0$, where

$$h(\mathbf{x}) = \begin{cases} \sum_i \dfrac{c_i(x_i)}{n x_i} & \text{if AC} \\ \sum_i \dfrac{c_i(x_i)}{X} & \text{if OR.} \end{cases}$$

Note that when $n = 1$, OR and AC regulation are the same, while they differ substantially (for nonsymmetric
particular, the OR constraint is a mark
constraint; $h(\mathbf{x}) = \Sigma\{[c_i(x_i)/x_i] \cdot (x_i/X)\}$

A simple two-firm example helps illu
straints. For convenience, consider the
facing constant average costs of produ
is, $c_i(x_i) = c_i x_i$, $i = 1, 2$, and $p(X) = a$
manipulation, the constraint can be writte

$$\frac{a}{b} - \frac{c_1 + c_2}{2\alpha b} \leq x_1 + x_2. \tag{1}$$

---

[2] Nonprice competition has been observed in some group-regulatory settings (see Douglas and Miller 1974). While the model I employ does not allow for product differentiation directly, the use of a price ceiling allows firms to adjust prices downward in an attempt to compete with each other. Thus, allowing for downward flexibility of prices acts as a surrogate for nonprice competition. I choose to use a homogeneous model to make clear welfare comparisons; the inclusion of product differentiation simply makes welfare comparisons mathematically difficult without adding new insights.

[3] If one defines a notion of an operating cost (such as short-run variable cost evaluated at the optimal fixed factor levels for output level x), one could express the OR constraint using this cost rather than $c_i(\,)$. This needlessly complicates the exposition.

Therefore, the constraint provides a simple lower bound on aggregate output and appears as in figure 1*a*. Since aggregate output is being restricted to be at least the value on the left-hand side of (1), the constraint acts to limit price to be no greater than $(c_1 + c_2)/2\alpha$.

If both firms have cost functions with fixed costs and constant marginal costs, that is,

$$c_i(x_i) = \begin{cases} c_i x_i + f_i & x_i > 0, \quad i = 1, 2, \\ 0 & x_i = 0, \end{cases}$$

then condition (1) becomes the following ($x_1 > 0$, $x_2 > 0$):

$$(c_1 + c_2 - 2\alpha a)x_1x_2 + f_1x_2 + f_2x_1 + 2\alpha b(x_1 + x_2)x_1x_2 \geqslant 0, \quad (2)$$

which is clearly not linear. An example of (2) is illustrated in figure 1*b*. Note that in this case the constraint actually acts more to restrict potential allocations of aggregate output to individual firms rather than actually restrict the price. While any price (or, alternatively, aggregate output level X) is regulatory feasible (i.e., $F \geqslant 0$), at some prices certain vectors of firm outputs are not regulatory feasible (those in the region $F < 0$). Thus, in this case, the constraint acts to restrict the distribution of output over firms in the industry.

In general I will assume that demand is downward sloping and concave, though weaker assumptions are possible. Thus, if $c_i(\cdot)$ is convex for each i, or if $c_i(x_i)/x_i$ is convex in the AC case, then $F(\mathbf{x}, \alpha)$ is a convex function. As Takayama (1969) noted in the Averch-Johnson analysis, the resulting constraint generally describes a nonconvex feasible region as, for example, in figure 1*b*.

I will be interested only in $\mathbf{x}$ such that $F(\mathbf{x}, \alpha) = 0$. Moreover, by requiring $F(\mathbf{x}, \alpha)$ to be a convex function I can satisfy the Arrow-Hurwicz-Uzawa (1961) constraint qualification and employ Kuhn-Tucker-Lagrange theory for this problem (see Takayama [1969], where this issue is first discussed for the A-J problem). Finally, I assume that the regulatory parameter is set by the regulator so that the constraint is binding on the industry. This will be important for the case $n \geqslant 2$, since this does not imply that each firm necessarily meets the constraint.

Regulatory Scenario

The regulatory scenario used in the paper is the following:

1. The regulator announces a procedure for regulation (i.e., a group regulatory constraint F) and any necessary parameter value (α).
2. The firms come to an agreement as to how to share the market

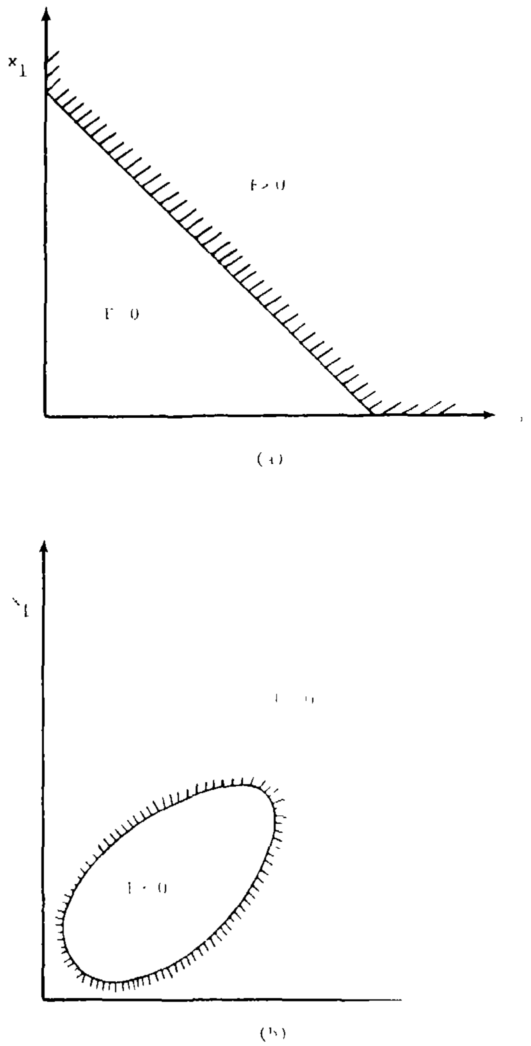


FIG. 1.—The average cost constraint (*a*) Constant average costs, linear demand (*b*) Constant marginal cost and a fixed cost, linear demand

($\mathbf{x} \in R^n_+$) and what price to propose, $p(X)$, subject to the regulatory constraint. The firms now provide the regulator with the price proposal and documentation to support the claim that the constraint is met.

3. The regulator accepts the industry's proposed price (since it satisfies the regulation).

4. The regulated price acts as a price ceiling; all firms use the ceiling price to guide their output decisions.

Step 2 in the sequence above is quite realistic. In general firms under regulation are often provided antitrust immunity, especially if the regulation calls for joint price proposals. Usually some sort of official institution (often called "rate bureaus") has evolved wherein the decisions required in step 2 (i.e., the price proposal) can be developed.

In the analysis I do not consider the game between the regulator and the industry (i.e., the proper determination of F and α). In the multifirm analysis, I will not allow firms to misrepresent or inflate their costs, though I will examine briefly what happens in the single-firm case when they can. Finally, no dynamic aspects are considered. All of these issues are potential avenues of future research but beyond the scope of the present paper.

The Single-Firm Industry

Consider a single-product monopoly firm that faces a known, downward-sloping inverse demand function $p(\cdot)$ and is subject to a cost function $c(\cdot)$. Since there is only one firm in the industry, the regulatory constraint is $[c(X)/X] - \alpha p(X) \geqslant 0$, where X is firm output and $\alpha \in [0, 1]$ is a parameter set by the regulator. Thus, the problem faced by the firm is to maximize profits subject to meeting the constraint imposed by the regulator:

$$\max p(X)X - c(X) \tag{3}$$

subject to $[c(X)/X] - \alpha p(X) \geqslant 0$.

Figure 2 illustrates the result of the regulatory constraint. Without regulation, optimal output is X^U at price p^U. The line αD provides the line of reference for regulated prices (i.e., α times the demand price). At X^U we see that αp^U exceeds AC. By expanding output to where AC $= \alpha D$ the firm now chooses X^R at price p^R. Thus the regulated price p^R is lower than the unregulated price p^U, presumably the intention of regulation.

Cherry (1978) and Moore (1978) have observed that such regulation can also generate A-J-like distortions under certain conditions. Both analyze the one-firm operating ratio case wherein costs are essentially determined by the level of labor employed (in both cases, the analysis is in the context of the trucking industry).[4] This is essentially the Averch-Johnson model with labor the focus of the regulatory constraint rather than capital (see Baumol and Klevorick [1970] for a detailed discussion of the symmetric case concerning capital bias).

---

[4] Other studies of single-firm operating ratio regulation include Nevel and Miklius (1968), Westfield (1974), and Braeutigam (1981).

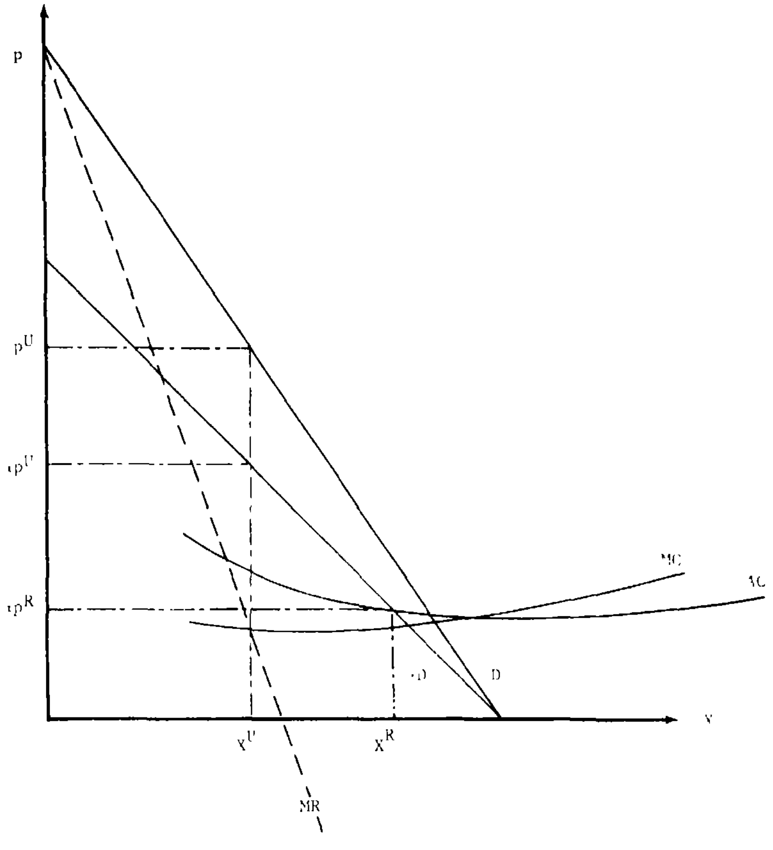


FIG. 2.—Single-firm

Thus, in this case, for any specified leve
a more labor-intensive input mix than w
level of output) in the unregulated environm
both find an "input bias" in the trucking industry that they attribute to the use of the operating ratio: *regulation acts to implicitly reduce the* real cost of labor, thereby inducing the firm to produce inefficiently (i.e., off its unregulated expansion path).

Note that in the discussion above, the firm is assumed to truthfully report its cost function to the regulator. In general this assumption may be false since regulators know only what firms report, as has been observed by Loeb and Magat (1979) and more recently by Baron and Myerson (1982). This can lead to the firm's choosing to inflate its cost

[5] As is well known, there is a large, conflicting empirical literature on the presence of A-J effects in public utilities (see, e.g., Courville 1974; Spann 1974, McKay 1976)

report to the regulator. Presumably, auditing by a regulator might detect outright lying across time, and thus the firm might actually engage in inefficient activities to drive costs up. Alchian and Kessel (1962) analyzed this possibility in the context of managers of regulated firms who engage in discriminatory hiring practices so as to capture any nonpecuniary benefits of associating with preferred (class, race, sex) employees.

That such actions not only are utility maximizing for some employers but can be profit maximizing for the regulated firm can be seen by the following example. Let $p(X) = a - bX$ and $c(X) = (c + \omega)X$, where c is the real marginal cost of production and ω is waste. Returning to (3), it is clear that maximizing profits subject to the regulatory constraint is accomplished by adjusting output to maximize revenue—or, more precisely, $(1 - \alpha)$ times revenue subject to the restriction that $c(X) \geqslant \alpha p(X)X$. If the firm can now also adjust costs, then profit maximization occurs at revenue maximization. This is accomplished by allowing cost to inflate until the average cost curve intersects αD at that output where $MR = 0$ (illustrated in fig. 3). Optimal waste is positive if true AC intersects αD at an output greater than X^* (where $MR = 0$). Thus the earlier results of lowered price and expanded output relative to the unregulated outcome are still true, though more moderate in degree.

Therefore, the ability to adjust costs (through either faulty reporting or wasteful, though potentially utility-enhancing, activities) leads to a breakdown in the efficacy and value of the proposed regulation. Note that the recognition that waste can be profitable under regulation, when juxtaposed to the Cherry/Moore observation of A-J effects, suggests that such biases (to the degree they create optimal waste levels) will not be unwelcome to managers of average cost/operating ratio regulated firms.

The Multifirm Problem

The problem with the foregoing analysis is that average cost or operating ratio regulation is normally applied to an industry or a subindustry derived by applying the regulation on a regional basis. The welfare results regarding the effect of regulation on price provided by the single-firm case (e.g., price lower than the monopoly price) do not extend to the multifirm case. This is because firms recognize the interdependence of decisions in the group; that is, firms not only consider market demand and their own technology but also consider actions taken by other firms. Moreover, the result of efficiency distortion (à la Averch and Johnson) is significantly modified. The nature of firm interdependency has implications for the existence and extent

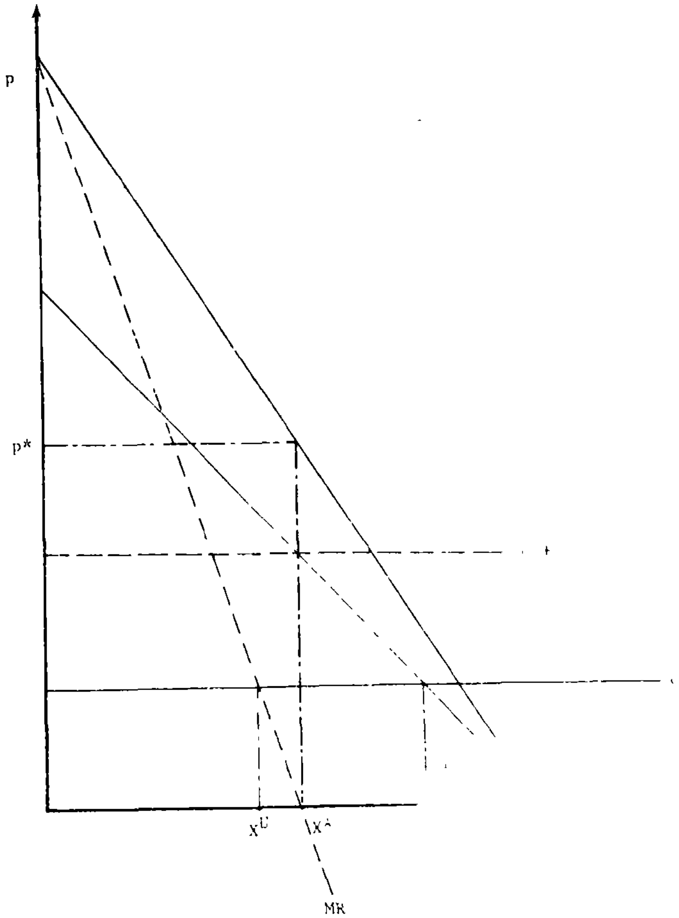


Fig. 3.—Cost inflation in the

of interfirm distortions that by definition are not observable in a single-firm analysis.

Thus, the multifirm problem must be approached as a problem of regulating an oligopoly. Of course, since there is no single way to treat oligopolies, the analysis of industry-wide regulation will depend on the assumptions made about the choice of "strategic" decision variables for the firm and the nature of firm interaction (the oligopoly solution). These points are examined in turn.

One of the interesting aspects of the industry-wide regulation problem is that the natural choice for decision variable for the firm is its output and not its price. Friedman (1977) makes just the opposite case for the (unregulated) oligopoly case in his discussion of Bertrand's objection to Cournot (p. 39): "There is an important objection to quantities as the decision variables of oligopolists, with a market price being determined by the total quantity produced. By what mechanism is the market price established? It is no problem imagining firms choosing prices and either producing output to order or producing for inventory. But what institutional arrangements accomplish price determination if the firms do not choose prices?" Under industry-wide regulation, however, such an institutional arrangement does exist: typically it is the industry that proposes a price (or price change) and the regulator passes on it. To make such a proposal the industry must pick a price that (1) satisfies the regulator's criteria and (2) guarantees a satisfactory "pie" for the members to divide. Since only one regulated price will be allowed, the crucial issue is the determination of the vector of outputs, and thus the shares of the pie. Thus, in the regulatory setting, output level is an appropriate decision variable, with aggregate output (in conjunction with the demand function) determining the price.

The allocation of market shares to member firms is a function of the organization of the industry, as represented by various oligopoly solutions (e.g., Cournot, Stackelberg, etc.). In what follows I employ a Stackelberg analysis involving a leader and followers. The leader will use the reaction functions of the followers, in conjunction with the regulatory constraint, to choose its output level and thus the distribution of output for the industry.

A Stackelberg Model of the Level and Distribution of Industry Output under the Regulatory Constraint

The use of industry-wide regulation affects firm and industry performance in three ways: via induced increases or decreases in the aggregate industry output level, via changes in the distribution of output by firms (i.e., market shares), and via affecting input level and mix choices. We consider output effects first and then consider input effects.

Output Level and Distribution

Consider an industry wherein a leader (in the Stackelberg sense) exists. It is important to emphasize that I view the leader as having power during step 2 of the regulatory scenario, that is, during the agreement step when firms must come up with a price proposal for the regulator. Once the ceiling price is set, the leader has no meaningful power: step 4 is not a sequential turn-taking process, wherein the leader can "go first." Both aspects of this assumption are reasonable. As indicated earlier, step 2 reflects the standard practice of allowing regulated firms to get together to propose price adjustments jointly to the regulator. On the other hand, since the application of this form of regulation is usually associated with the existence of a moderate-to-large number of firms, any leader at step 2 is unlikely to have much power at step 4, especially since a ceiling price has been set.

We represent this by examining the agreement formation (step 2 in the scenario) as a two-firm industry, with the leader as firm 1. The follower(s) (represented by firm 2) act in a Cournot manner, taking what the leader does as given. This would translate, in the usual (i.e., unregulated) analysis, into the leader taking the follower's reaction function, $\phi(x_1)$, as given in his profit maximization. This would then guarantee that an optimal solution to the leader's problem would provide an industry equilibrium.

Under regulation, however, the leader must incorporate the fact that in step 4 the follower takes both the leader's output *and* the ceiling price (once established) as given. Thus, in step 4 of the regulatory scenario, firm 2 solves the following problem: $\max_{x_2} \Pi^2(x_1, x_2)$ subject to $p(x_1, x_2) \leqslant \bar{p}$, where x_1 is the leader's output, $\bar{p}$ is the established ceiling price, and $\Pi^2(x_1, x_2) = p(x_1 + x_2)x_2 - c_2(x_2)$ is the follower's profit, which I assume is concave in x_2. Let $\Psi(x_1, \bar{p})$ be the reaction function for firm 2 (see App for details). This function specifies firm 2's optimal level of output given the leader's output choice (in step 2) and the price ceiling

Knowing $\Psi(x_1, \bar{p})$, the leader can pro
bution (i.e., a vector **x**) and price $p(\Sigma x_i$
scenario that will be an equilibrium in s
lem at step 2 is to pick a regulatory feas
a price ceiling $\bar{p}$ so that its profits are n
the following problem, denoted the lea

$$\max_{x_1, \bar{p}} \Pi^1[x_1, \Psi(x_1, p)]$$

subject to $p[x_1 + \Psi(x_1, \bar{p})] \leqslant \bar{p}$ and $F[x_1, \Psi(x_1, \bar{p}), \alpha] \geqslant 0$, where the first constraint assures that aggregate output will satisfy the price ceiling to be proposed and the second constraint requires that aggregate output and the price ceiling satisfy the regulatory constraint F

As shown in the Appendix, the solution to the leader's problem provides the following welfare result concerning the relationship of aggregate regulated output (X^R) to aggregate unregulated output (X^U):

$$X^R \lesseqgtr X^U \quad \text{if and only if} \quad (F_1 + F_2\phi')|_{\mathbf{x}=\mathbf{x}^R} \lesseqgtr 0, \qquad (4)$$

where $F_i \equiv \partial F/\partial x_i$. Condition (4) states that the question of whether or not regulation acts to produce equilibria that are welfare impairing (i.e., aggregate output reducing) is a function of the relationship between the marginal rates of output substitution along F and ϕ:

$$X^R \lesseqgtr X^U \Longleftrightarrow \left.\frac{dx_2}{dx_1}\right|_{\phi} = \phi' \lesseqgtr -\frac{F_1}{F_2} = \left.\frac{dx_2}{dx_1}\right|_{F}, \tag{5}$$

assuming $F_2 > 0$.

As a point of reference, figure 4*a* illustrates the unregulated Stackelberg solution $\mathbf{x}^U$ where a profit contour for firm 1 is just tangent to firm 2's reaction function ϕ. Figures 4*b* and 4*c* illustrate condition (4). In figure 4*b*, the leader maximizes profits by locating the industry at $\mathbf{x}^R$, which results in a price $\bar{p} = p(x_1^R + x_2^R)$. This price ceiling, in turn, induces an output level set that consists of the pairs of outputs on or above the line denoted $X(\bar{p})$. Note that, in the case illustrated, $\mathbf{x}^R$ is an equilibrium solution since firm 1 cannot increase its profits without raising the price, which is not possible. Figure 4*b* illustrates an example wherein welfare has increased in the sense that $X^R > X^U$ and $\bar{p} = p(X^R) < p(X^U)$.

As an example, consider the case wherein both firms have constant average costs: $c_i(x_i) = c_i x_i$, $i = 1, 2$. If the industry faces AC regulation then it is straightforward to show that $X^R > X^U$ under average cost regulation with constant average costs (note that in this case the regulatory constraint is as shown in fig. 1*a*). Hence, regulation "works" in the sense that output is increased and price reduced relative to the unconstrained solution.

Figure 4*c* illustrates just the opposite. Here $X^R < X^U$ and $\bar{p} > p(X^U)$. Note that any attempt (after the price ceiling is set) by firm 1 to increase its profits involves movement along ϕ above $\mathbf{x}^R$; this corresponds to setting price below the ceiling and lowering firm 2 profits. Firm 1 will not do this since (1) such an action would violate the regulatory constraint and (2) it would concomitantly require the other firms to cut their outputs, which they have no incentive to do. All firms simply see the price ceiling and react to it. The followers from step 2 take the leader's proposed output (as well as all the other followers' outputs) and end up at their agreed output levels. Knowing this, the leader also now acts in a Cournot manner by taking the price ceiling and other outputs as given and maximizing.

This is a particularly interesting case since x_2^R has unambiguously increased and x_1^R has unambiguously decreased relative to the unregulated equilibrium $\mathbf{x}^U$. To the degree that firm 2 has higher costs than firm 1, the regulatory restriction results in lower, and less efficiently produced, output.

As an example, consider again the situation wherein both firms face

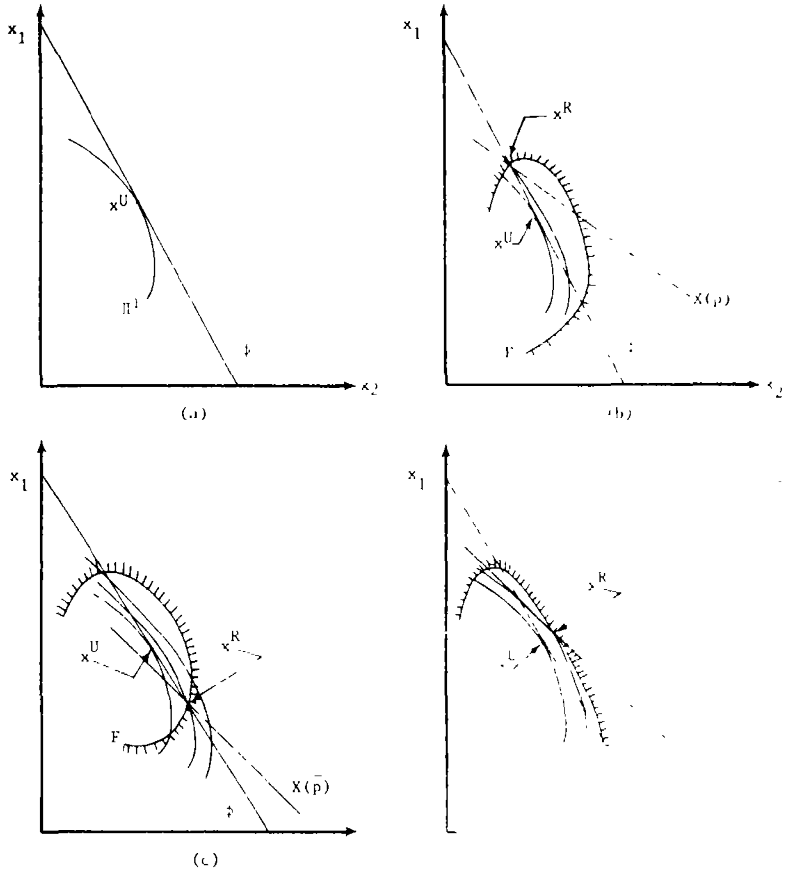


Fig. 4.—Stackelberg solutions. (*a*) Unregulated tion outcome (i) (*c*) Regulated, intersection outcome.

constant average costs of production. Let $p(X) = a$ [illegible] Furthermore, we consider the OR constraint to be the regulation in question. The unconstrained solution (i.e., ignoring the OR constraint) is $x_1^U = (a + c_2 - 2c_1)/2b$. By the assumption above, $x_1^U > 0$. Employing the first-order conditions (see [A1] and [A2] in the App.) for the leader in the regulated situation, we find that

$$\alpha > \frac{2c_1 + c_2}{a + c_2} \Rightarrow x_1^R < x_1^U \tag{6}$$

Again, since $a/2 > c_1$, we know that the ratio $(2c_1 + c_2)/(a + c_2)$ is a proper fraction. Thus if the left side of (6) is met the regulation will encourage the industry to produce less than would have otherwise occurred. Condition (6) is a sufficient condition, and weaker (and more complex) conditions exist.[6] For smaller and smaller values of α, a smaller and smaller value of c_1 is required by the condition. This reflects the fact that firm 1 could choose to face the constraint alone (by pricing so as to eliminate firm 2), but does not if profits are sufficiently high under market sharing. (A low value of α is associated with high profits in both the market-sharing and market-monopolizing states.)

Thus, if firm 1's costs are low enough (even under constant marginal costs for both firms), it has an incentive to share the market with a high cost firm. If c_2 is above the monopoly price that firm 1 would charge in an unregulated environment, then since $x_2^R > 0$, it must be that the regulated price is above the monopoly price. In other words, one can easily construct examples wherein not only is $X^R < X^U$, but X^U is the monopoly output and thus $p(X^R)$ is higher than the monopoly price. Society would be better off with an unregulated monopolist rather than the regulated industry. Moreover, since c_2 exceeds c_1, then the reduction in output is also accompanied by a shift in the distribution of output, with the higher-cost firm increasing its share of industry output.

Both of the examples above involve structurally competitive examples wherein firms have constant average costs of production. While examples of such regulated industries are far from unknown (e.g., trucking, airlines), presumably well-intentioned regulators know to avoid regulating such industries in the future and thus will turn their attention to industries that are not structurally competitive.

Unfortunately, both AC and OR regulation continue to have bad properties when applied to industries that are not structurally competitive. Specifically, let firm 2 be subject to constant returns to scale and firm 1 have economies of scale at the constrained solution such that

$$\frac{c_1'(x_1^R)}{x_1^R} - \frac{c_1(x_1^R)}{(x_1^R)^2} - 2\alpha p'(X^R) < 0. \tag{7}$$

Then $X^R < X^U$. This holds since under the condition above, $-F_1/F_2$ is positive and thus, from (4) above, $X^R < X^U$. Condition (7) reflects a situation wherein firm 1 has significant economies of scale at the regulated output vector $\mathbf{x}^R$. Note that this would presumably be a firm that, in an unregulated environment, might be a prototypical natural monopoly. However, since $X^R < X^U$ in the regulated equilibrium, this means that regulation has resulted in market sharing by the leader with the follower.

---

[6] Note that the existence of such examples is always assured, independent of the relationship between c_1 and c_2. This holds since, for any $\alpha \in (0, 1)$ one can always find a number V such that if $c_1 < a/V$ then $\alpha > (2c_2 + c_2)/(a + c_2)$ Note that while the actual value of V depends on c_1 and c_2, its existence does not

Again, this sort of result holds in the operating ratio case, too. For example, let $p(X) = 33 - (13/14)X$, $c_1(x_1) = 8\sqrt{x_1}$, $c_2(x_2) = 20x_2$. The unregulated solution involves firm 1's acting as a monopolist and producing approximately 17.25 units. If the operating ratio is set at .52, the resulting Stackelberg solution is $x_1 = 4$, $x_2 = 5$ for a total output of 9 units. Thus, regulation severely restricted output to essentially half the monopoly output.[7] Moreover, it should again be noted that the market share of the high cost firm has increased.

This seemingly counterintuitive, clearly noncompetitive behavior is understandable if we observe that a firm with rapidly falling average costs could reasonably prefer to share a market with a higher-cost competitor, and thereby keep the regulated price high and reap some profits, rather than be forced out along its average cost function with a significantly lowered price. Thus, if the leader is an efficient firm, it has an incentive to distribute output (and propose a price ceiling) so as to keep an inefficient firm operating, thereby "padding the average."

Finally, figure 4*d* illustrates the condition wherein the optimal solution occurs where the isoprofit contour is tangent to the regulatory constraint contour (see the App., eq. [A1]). In this particular situation $x_2^R > \phi(x_1^R)$; that is, follower output under regu
would be without regulation for the g
reflecting the fact that the ceiling price
choices. Moreover, note that there is no
choices and improve, since once the ceil
higher price than $\bar{p}$. Finally, we observe
may involve increases or decreases in w

Input Levels and Mix Effects

The Stackelberg model also provides implications about distortions in the input markets for the industry. Since the leader must observe the constraint, then to the degree the constraint emphasizes any particular factor of production (as Moore and Cherry suggest about the operating ratio), an A-J-like bias will occur in the leader's input choices. However, the followers will be free of an input-mix bias since they never directly need observe the regulatory constraint F.

---

[7] While such a result would seem rather extreme, Robert Forsythe and I used the parameters above in an experiment involving two students acting as the two duopolists. They both knew the demand schedule and their own cost schedule, they were paid the profits they made. After a few rounds of cheating on agreements, followed by a threat of output dumping by firm 2 should firm 1 cheat again, followed by some attempts at collusion, they converged to (essentially) the predicted outcome ($x_1 = 3$, $x_2 = 5$). It is worth noting that, to our knowedge, Stackelberg equilibria have never before been generated in experiments. It should also be noted that previous experiments have been with unregulated oligopolies

As illustrated above, all firms are generally affected with respect to output-level setting. Of course, to the degree that output is shifted to the leader, then to that degree the A-J-like distortion becomes more significant. On the other hand, to the degree that the market share of follower firms grows, then the input-bias distortion suffered by the leader becomes less significant.

Finally, let us consider once again the wasteful monopolist examined earlier. Nowhere in the multifirm analysis have we directly allowed for cost inflation. The single, wasteful monopolist indicates why this is not a serious liability for the analysis. Note that another way to interpret the wasteful monopolist is as a two-firm model (with truthful revelation of costs) with a nonwasteful monopolist as a Stackelberg leader and the follower providing the necessary cost increase. Thus the incentive to pad by setting price high enough to maintain inefficient firms subsumes the case of misrepresentation of efficient production costs. To put this another way, allowing for cost inflation would simply further reinforce the results found under truthful revelation.

This is not to imply that firms will not choose to misrepresent their costs. In the Stackelberg model, however, only the leader has an incentive to misrepresent costs since only the leader faces the F-constraint. Followers have no incentive to be inefficient since they only face the output of the leader and the ceiling price restriction. Thus, only the leader has an incentive to waste and only the leader may suffer from an A-J-like input bias. Of course, since these distortions affect the level of leader output, industry output is also affected.

Summary and Conclusions

Historically, the motivation for the use of industry-wide regulation (as opposed to case-by-case review) reflected an attempt to reduce two costs of regulation: the administrative burden associated with dealing with a large number of firms, and lack of information about the likely costs and profitability of individual firms. The first cost was eased by providing across-the-board price adjustments based on aggregate performance, the second by reference to aggregate experience. The rationale for the Reagan administration's planned medicare payment scheme based on national average hospital costs is that such a procedure is a "deregulation" of the system. This rationale especially reflects the first cost concern.

The substitution of industry-based controls for firm-level controls turns out to trade one set of regulation-induced costs for another. Presumably, the beneficiaries of regulation, and the regulator, find this substitution advantageous (see Peltzman 1976). Especially to the degree that a price-leader form of organization occurs, and with it padding the average, the benefits of regulation are conferred on inefficient firms (by being allowed to operate) and efficient firms (in the form of protected profits) by using industry-wide controls. This occurs simultaneously with the perception that, on average, the industry is operating at a "fair" relationship between price and cost.

Of course, free entry can undo this result, which may account for the existence of entry controls in such settings. A classic example is the motor carrier industry under ICC regulation during the forties, fifties, and sixties. As Stigler (1971) shows (see his fig. 1), in the face of steadily rising applications for entry into the industry, the number of actual licensed carriers fell. This occurred precisely during the period wherein the ICC employed a 93 percent operating ratio as its sole criterion for decisions on rate increases (see Dobesh 1973). It was also during this period that significant concentration developed in the industry (see LaMond 1980) in spite of no appreciable scale economies (see Spady and Friedlaender 1978). This is suggestive (though certainly not conclusive) evidence of the trucking industry as a real-world example of the discussion outlined above.

It is also worth noting that entry controls [illegible] convenient way for firms in the industry [illegible] entrants allowed in. This holds since [illegible] cumbents to object to potential entrants [illegible] necessity" basis. Since powerful firms [illegible] efficient entrants (who are most likely [illegible] and least likely to object to inefficient [illegible] able to or not choose to compete with [illegible] is likely to screen out low-cost entrants [illegible] trants.

As indicated in the Introduction, a number of industries besides trucking have been subjected to industry-wide price regulation. Moreover, variations on this theme have been used at the firm level, too. For example, under the Carter administration the Council on Wage and Price Stability regulated individual firms by allowing them to average over corporate divisions. The results of my analysis would suggest that, to the degree such regulation was restrictive, some firms may have purposely kept inefficient divisions alive rather than divest. Again, the administrative advantages of such regulatory procedures make them a pervasive tool of control.

Thus, the use of industry-wide regulation based on aggregate performance measures generates a number of distortions in industry structure, conduct, and performance. In general such regulation encourages the setting of prices high enough to maintain inefficient firms, and inefficient firm operation due to incentives to inflate costs and to A-J effects. The distribution of output in the affected industry is thus quite different from that which would occur without regulation, with a greater share of industry output being provided by inefficient firms than would otherwise occur.

Appendix

The Stackelberg follower solves the following problem:

$$\max \pi^2(x_1, x_2)$$

subject to $p(x_1 + x_2) \leq \bar{p}$ given $\bar{p}$ and x_1. If the profit function is concave in x_2 and the constraint is nonbinding, then the firm's optimal choice of output level can be written as $x_2 = \phi(x_1)$, where ϕ is a monotonically nonincreasing function of firm 1 output. Moreover, concavity guarantees that the slope of the reaction function ϕ is a negative fraction, that is, $\phi' \in (-1, 0)$.

If the constraint is binding, for given x_1 and $\bar{p}$, then optimal x_2 is simply that output such that $p(x_1 + x_2) = \bar{p}$. Thus the resulting regulated firm reaction function $\Psi(x_1, \bar{p})$ is

$$\Psi(x_1, \bar{p}) = \begin{cases} \phi(x_1) & p(x_1 + x_2) < \bar{p} \\ X(\bar{p}) - x_1 & \text{otherwise,} \end{cases}$$

where $X(p)$ is the demand function. An example of Ψ is illustrated in figure A1, where for convenience x_1 is placed on the vertical axis. A kink occurs where ϕ intersects $X(\bar{p})$. For lower $\bar{p}$ (large $X[\bar{p}]$) the kink occurs higher up ϕ, closer to the x_1 axis. Since $X(\bar{p})$ is the line $x_1 + x_2 = p^{-1}(\bar{p})$, it has a slope of -1. Thus, letting subscripts on Ψ denote partial derivatives (or subgradients as necessary):

$$\Psi_1(x_1, \bar{p}) \begin{cases} = \phi' & p(x_1 + x_2) < \bar{p},\ x_2 = \phi(x_1) \\ \in [-1, \phi'] & p(x_1 + x_2) = \bar{p},\ x_2 = \phi(x_1) \\ = -1 & \text{otherwise} \end{cases}$$

Thus, while at the kink there is no derivative, the change in optimal x_2 for a change in x_1 (given $\bar{p}$) is between -1 and the slope of the unregulated reaction curve.

Changes in $\bar{p}$ are directly computed from the definition of Ψ:

$$\Psi_2(x_1, \bar{p}) = \begin{cases} 0 & p(x_1 + x_2) < \bar{p} \\ \dfrac{1}{p'} & \text{otherwise,} \end{cases}$$

since in the first case ϕ is not a function of $\bar{p}$ while in the second case $X(p)$ and $p(X)$ are inverses.

The Stackelberg leader solves the following problem:

$$\max_{x_1 \bar{p}} \Pi^1[x_1, \Psi(x_1, \bar{p})] \qquad \text{(LP)}$$

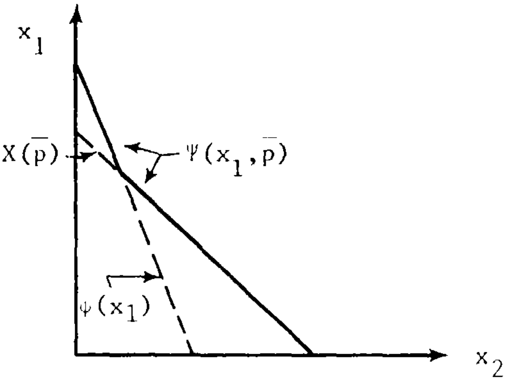


FIG. A1 –Regulated firm reaction function

subject to $p[x_1 + \Psi(x_1, \bar{p})] \leqslant \bar{p}$ and $F[x_1, \Psi(x_1, \bar{p}), \alpha] \geqslant 0$.

Let μ and λ be nonnegative Kuhn-Tucker multipliers on the first and second constraint and let $\Pi_i^1 \equiv \partial \Pi^1(x_1, x_2)/\partial x_i$ and $F_i \equiv \partial F(x_1, x_2, \alpha)/\partial x_i$. The first-order conditions for an interior solution are as follows:

$$x_1\colon \Pi_1^1 + \Pi_2^1\Psi_1 + \lambda(F_1 + F_2\Psi_1) - \mu p'(1 + \Psi_1) = 0 \tag{A1}$$

$$\bar{p}\colon \Pi_2^1\Psi_2 + \lambda F_2\Psi_2 + \mu(1 - p'\Psi_2) = 0 \tag{A2}$$

Examining (A1) and (A2), we see that there are two possibilities of interest. If $p(x_1 + x_2) < \bar{p}$ then $\mu = 0$, $\Psi_2(x_1, \bar{p}) = 0$, and $\Psi_1(x_1, \bar{p}) = \phi'$. Thus (A1) and (A2) reduce to the following.

$$\Pi_1^1 + \Pi_2^1\phi' = -\lambda(F_1 + F \quad \tag{A3}$$

Alternatively, if $\mu > 0$ (and thus we
then $\Psi_2(x_1, \bar{p}) = 1/p'$ and $\Psi_1(x_1, \bar{p}) =$
condition:

$$\frac{\Pi_1^1}{\Pi_2^1} =$$

To understand the conditions above,
tory constraint F admitted the unconstrained
tion (A3) reduces to the standard Stackelberg [illegible]
curve for firm 1 is tangent to the unconstrained firm 2 reaction function ϕ.
Let us assume now that α is such that $F(\mathbf{x}^{\prime}, \alpha) < 0$ and that $\lambda > 0$. Condition (A3) can be developed to provide a welfare implication. To see this, note that aggregate output X is simply $x_1 + x_2 = x_1 + \phi(x_1)$ and therefore (since ϕ' is a negative fraction):

$$\frac{dX}{dx_1} = 1 + \phi'(\mathbf{x}) > 0$$

In other words, increasing the leader's output level, x_1, is synonymous with increasing industry output, while decreasing x_1 yields a reduction in X. Therefore we have a ready indicator of the welfare implications of regulation increases in leader output vis-à-vis the unregulated equilibrium imply increases in welfare

The term on the left of (A3) is the total derivative of Π^1 with respect to x_1 along ϕ, $(d\Pi^1/dx_1)$, which is set to zero in the unconstrained case. If the right-hand side of (A3) is positive, then marginal leader profit is positive and thus the constraint is restricting the leader from producing as much output as would be produced in the unconstrained equilibrium, $\mathbf{x}^U$. On the other hand, if the right side of (A3) is negative, then marginal leader profit is negative and the effect of the regulation is to force the leader to produce more than x_1^U. Since aggregate output is monotonic in the leader output, we have the following result when the F constraint is restrictive (i.e., $\lambda > 0$):

$$X^R \gtreqless X^U \text{ if and only if } (F_1 + F_2\phi')|_{\mathbf{x}=\mathbf{x}^R} \gtreqless 0. \qquad \text{(A5)}$$

Finally, (A4) reflects the situation wherein the solution to (LP) involves a tangency between an isoprofit contour and the regulatory constraint F.

References

Alchian, Armen A., and Kessel, Reuben A. "Competition, Monopoly, and the Pursuit of Pecuniary Gain." In *Aspects of Labor Economics*. Princeton, N.J.: *Princeton Univ. Press (for N.B.E.R.), 1962.*

Arrow, Kenneth J.; Hurwicz, Leonid; and Uzawa, Hirofumi. "Constraint Qualifications in Maximization Problems." *Naval Res. Logistics Q.* 8 (June 1961): 175–91.

Baron, David P., and Myerson, Roger B. "Regulating a Monopolist with Unknown Costs." *Econometrica* 50 (July 1982): 911–30.

Baumol, William J., and Klevorick, Alvin K. "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion." *Bell J. Econ. and Management Sci.* 1 (Autumn 1970): 162–90.

Braeutigam, Ronald R. "The Regulation of Multiproduct Enterprises by Rate of Return, Markup, and Operating Ratio." In *Research in Law and Economics*, vol. 3, edited by Richard Zerbe. Greenwich, Conn.: JAI, 1981.

Cherry, Russell. "The Operating Ratio Effect and Regulated Motor Carriers." In *Proceedings of a Workshop on Motor Carrier Economic Regulation*. Washington: Nat. Academy Sci., 1978.

Courville, Léon. "Regulation and Efficiency in the Electric Utility Industry." *Bell J. Econ. and Management Sci* 5 (Spring 1974): 53–74.

Dobesh, Larry J. "Earnings Control Standards for Regulated Motor Carriers." Ph.D. dissertation, Washington State Univ , 1973.

Douglas, George W., and Miller, James C., III. "The CAB's Domestic Passenger Fare Investigation." *Bell J. Econ. and Management Sci.* 5 (Spring 1974): 205–22

Friedman, James. *Oligopoly and the Theory of Games.* Amsterdam: North-Holland, 1977.

Harris, Seymour E. *Price and Related Controls in the United States.* New York: McGraw-Hill, 1945.

Hirshleifer, Jack. *Price Theory and Applications.* Englewood Cliffs, N.J.: Prentice-Hall, 1976.

Ippolito, Richard A. "The Effects of Price Regulation in the Automobile Insurance Industry." *J. Law and Econ.* 22 (April 1979): 55–89.

Joskow, Paul L. "Cartels, Competition and Regulation in the Property-

Liability Insurance Industry." *Bell J. Econ. and Management Sci.* 4 (Autumn 1973): 375–427.

Kitch, Edmund W.; Isaacson, Marc; and Kasper, Daniel. "The Regulation of Taxicabs in Chicago." *J. Law and Econ.* 14 (October 1971): 285–350.

LaMond, Annette M. *Competition in the General Freight Motor-Carrier Industry* Lexington, Mass.: Heath, 1980.

Locklin, D. Philip. *Economics of Transportation* 7th ed. Homewood, Ill.: Irwin, 1972.

Loeb, Martin, and Magat, Wesley A. "A Decentralized Method for Utility Regulation." *J. Law and Econ.* 22 (October 1979): 399–404.

MacAvoy, Paul W. *The Regulated Industries and the Economy.* New York: Norton, 1979.

McKay, Derek. "Has the A-J Effect Been Empirically Verified?" Social Science Working Paper no. 132. Pasadena: California Inst. Tech., 1976.

Moore, Thomas G. "The Beneficiaries of Trucking Regulation." *J Law and Econ.* 21 (October 1978): 327–43

Nevel, R., and Miklius, Walter. "The Operating Ratio as a Regulatory Standard." *Transportation J.* 8 (Winter 1968): 15–18.

Peltzman, Sam. "Toward a More General Theory of Regulation." *J Law and Econ.* 19 (August 1976): 211–40

Spady, Richard H., and Friedlaender, Ann F. "Hedonic Cost Functions for the Regulated Trucking Industry." *Bell J. Econ.* 9 (Spring 1978): 159–79

Spann, Robert M. "Rate of Return Regulation and Efficiency in Production. An Empirical Test of the Averch-Johnson Thesis." *Bell J Econ and Management Sci.* 5 (Spring 1974): 38–52.

Stigler, George J. "The Theory of Economic Regulation" *Bell J Econ. and Management Sci.* 2 (Spring 1971): 3–21.

Takayama, Akira. "Behavior of the Firm under Regulatory Constraint." *A.E.R.* 59 (June 1969). 255–60.

Westfield, Fred "The Operating Ratio as Regul [illegible] Paper presented at North American Meeting of [illegible] cisco, December 1974.

On Measuring Natural Resource Scarcity

Robert Halvorsen
University of Washington

Tim R. Smith
Federal Reserve Bank of Kansas City

Conclusions concerning trends in natural resource scarcity may depend critically on the choice of scarcity index. Unfortunately, the prevalence of vertical integration in natural resource industries has hindered the use of some otherwise desirable scarcity measures. In this paper duality theory is used to derive an econometric procedure for estimating one such measure, the shadow price of the resource in situ. Empirical results for the Canadian metal mining industry indicate that resource scarcity as measured by this shadow price has decreased substantially over time.

I. Introduction

Natural resource scarcity and its implications for economic growth have become important areas of economic research in recent years. One issue that has received considerable attention is how best to measure changes in the economic scarcity of exhaustible natural resources.[1] Candidate measures include the prices of natural resource

---

Research for this paper was partially supported by an Alfred P. Sloan Foundation grant to the University of Washington for research in natural resource economics. We are grateful to Lee E. Edlefsen, V. Kerry Smith, and Joseph E. Swierzbinski for comments on an earlier version of this paper. The views and opinions expressed herein are solely those of the authors and do not necessarily represent those of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

[1] The pioneering study of the measurement of natural resource scarcity is Barnett and Morse (1963). Smith (1979*b*) contains a number of useful papers reevaluating the Barnett and Morse work. Other recent studies include Brown and Field (1978), Pin-

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0013$01.50

products at various stages of processing, the price of the resource in situ (also referred to as the resource's rent, royalty, or net price), and the unit cost of extraction.

Increasing physical scarcity may be associated with either increases or decreases over time in any or all of the proposed measures. Because they respond to different sets of economic and technological influences, there is no reason to expect the magnitude, or even the direction, of scarcity change to be the same for alternative measures. Therefore, no one measure of natural resource scarcity can be said to be the most appropriate in general on theoretical grounds. Instead the choice of a scarcity measure in a particular context should be based on an evaluation of which economic and technological effects on scarcity are relevant for the question at hand. At the same time, it is critical that the preferred scarcity measure be estimated directly, since the use of a proxy measure might lead to highly misleading results.[2]

In practice, the measurement of natural resource scarcity has been severely constrained by data considerations. Owing in large part to the prevalence of vertical integration in natural resource industries (Manthy 1978), data on the prices of resources in situ, unit extraction costs, and the prices of extracted but unprocessed resources have been very difficult to obtain. As a result, there has been a tendency to rely on the prices of processed resource products in measuring trends in natural resource scarcity.[3]

In this paper we show how duality theory can be used to derive estimates of the competitive price of the resource in situ from data on
vertically integrated natural resource [illegible]
illustrated with data for the Canadi[illegible]
empirical results indicate that the sl[illegible]
has fallen substantially from 1956 to [illegible]
cessed ore has increased slightly.

The derivation of the economet[illegible]

---

dyck (1978), Smith (1978, 1979*a*), Johnson, Bell, and [illegible]
Fisher (1982), and Slade (1982) Smith (1980) and Fisher (1981) provide [illegible] overviews of the literature in this area

[2] For further discussion see Brown and Field (1978), Smith (1978, 1980), Hanson (1980), Fisher (1981), and Halvorsen and Smith (1984)

[3] See, e.g , Barnett and Morse (1963), Smith (1979*a*), and Slade (1982) Barnett and Morse (1963) also attempt to measure unit extraction costs, but their estimates include costs of some processing activities and do not reflect changes in input prices Their cost measures have been updated by Johnson et al. (1980). Brown and Field (1978) report the limited available data on the prices of resources in situ They suggest, as does Fisher (1981), the use of marginal exploration cost as an indirect measure of the scarcity rent, but Devarajan and Fisher (1982) show that marginal exploration cost need not be equated to the price of the resource in situ given uncertainty in exploration Moroney and Trapani (1981) simulate the effects on the output price of mineral-intensive industries of changes in the prices of capital, labor, and minerals

section. The specification of the model and the empirical results are discussed in Section III.

II. The Model

The production function for a vertically integrated natural resource firm is assumed to be

$$Q = Q(\mathbf{X}, Z, T), \tag{1}$$

where Q is the quantity of final output, $\mathbf{X}$ is a vector of reproducible inputs, Z is the stock of the exhaustible natural resource, and T (time) allows for the effects of technological change. Assuming that the quantities of inputs used in extraction are separable from those used in processing, we can rewrite equation (1) as

$$Q = Q[\mathbf{X}^P, T, N(\mathbf{X}^E, Z, T)], \tag{2}$$

where the superscripts P and E represent processing and extraction activities, respectively, and N is the output of the extraction subproduction function, that is, the quantity of ore extracted.

The firm is assumed to maximize the wealth obtainable from its stock of the natural resource given input and output prices, the technological conditions governing extraction and processing, and the constraints

$$Z_t = Z_0 - \int_0^t N_\tau d\tau \geqslant 0,$$

$$\dot{Z} \equiv \frac{dZ}{dt} = -N_t.$$

The current-valued Hamiltonian for the firm's wealth maximization problem is

$$H = P_Q Q(\mathbf{X}^P, N, T) - \sum_i P_i X_i^P - C(N, \mathbf{P}_\mathbf{X}^E, Z, T) - \mu N,$$

$$i = 1, \ldots, n,$$

where μ is a costate variable, r is the rate of discount, P_Q is the price of output, $\mathbf{P_X}$ is a vector of input prices, and $C(\cdot)$ is the minimal total cost function dual to the extraction subproduction function. If a competitive market existed for the natural resource in situ, the market price in period t would be equal to μ_t because this represents the marginal opportunity cost to the firm of one unit of stock.

Our procedure for estimating μ is based on the reproducible cost function dual to the production function for final output, equation (2). The solution to the firm's wealth-maximizing problem gives the optimal quantities of N to extract and of Q to produce in each period. The reproducible cost function can be derived by considering the auxiliary problem of minimizing the total cost of reproducible inputs in each period given Z and the optimal quantities of Q and N in each period.[4]

With the time subscripts omitted, the Lagrangian for the constrained cost minimization problem is

$$\begin{aligned}\mathcal{L} = \sum_i P_i(X_i^P + X_i^E) + \theta\{Q - Q[\mathbf{X}^P, T, N(\mathbf{X}^E, Z, T)]\}\\ + \delta[N - N(\mathbf{X}^E, Z, T)].\end{aligned} \tag{3}$$

The values of Q and N are taken as set at their wealth-maximizing levels. The first-order conditions for the cost-minimizing quantities of reproducible inputs in processing and extraction are, respectively,

$$P_i = \theta \frac{\partial Q}{\partial X_i^P}, \tag{4}$$

$$P_i = \delta \frac{\partial N}{\partial X_i^E}. \tag{5}$$

The solution of this cost minimization problem yields the reproducible cost function,

$$CR = CR(Q, \mathbf{P_X}, N, Z, T). \tag{6}$$

By the envelope theorem (Silberberg 1978

$$\frac{\partial CR}{\partial N} = \frac{\partial \mathcal{L}}{\partial N} =$$

The right-hand side of (7) can be
solution of the total cost minimizatio
The Lagrangian is

$$\begin{aligned}\tilde{\mathcal{L}} = \sum_i P_i(X_i^P + X_i^E) + \mu N(\mathbf{X}\ , Z, T)\\ + \tilde{\theta}\{Q - Q[\mathbf{X}^P, T, N(\mathbf{X}^E, Z, T)]\},\end{aligned} \tag{8}$$

---

[4] Of course, firms will not explicitly solve the restricted cost minimization problem considered here but instead will solve simultaneously for the wealth-maximizing quantities of Q and N together with the optimal quantities of reproducible inputs. However, the optimal quantities of reproducible inputs given by the solution to the restricted cost minimization problem will be identical to the quantities implied by the more general wealth-maximizing problem (see Lau 1976). Note that the cost minimization problem is consistent with the intertemporal control problem since the Hamiltonian essentially summarizes an infinite series of static optimization problems

where μ is the (unobservable) shadow price of the ore in situ. The first-order conditions for the reproducible inputs are

$$P_i = \bar{\theta} \frac{\partial Q}{\partial X_i^P}, \tag{9}$$

$$P_i = \left(-\mu + \bar{\theta} \frac{\partial Q}{\partial N}\right) \frac{\partial N}{\partial X_i^k}. \tag{10}$$

The solution values, $\mathbf{X}^k$ and $\mathbf{X}^P$, for this minimization problem will be identical to those derived for the restricted problem with N set equal to its wealth-maximizing level.

From (4) and (9),

$$\theta = \bar{\theta}. \tag{11}$$

From (5), (10), and (11),

$$\delta = -\mu + \theta \frac{\partial Q}{\partial N}. \tag{12}$$

Substituting from (12) in (7) we get

$$\frac{\partial CR}{\partial N} = -\mu. \tag{13}$$

Thus estimates of the shadow price of the ore in situ can be obtained from data for a vertically integrated natural resource industry by estimating its reproducible cost function and then differentiating the cost function with respect to N.[5] The econometric model used to implement this procedure for the Canadian metal mining industry is described in the following section.

III. Econometric Specification and Empirical Results

The output of the Canadian metal mining industry is specified to be a function of metallic ore, three reproducible inputs, capital, K, labor, L, and energy, E, and time, T.[6] We use the translog functional form for the cost function,

---

[5] The result that the partial derivative of a restricted cost function with respect to a quantity variable is equal to the negative of its shadow price was first noted by Hotelling (1932) and is generally referred to as Hotelling's lemma

[6] Data restrictions preclude the inclusion of the stock of the resource, Z, in the econometric model. The omission of Z will not bias the estimates of the parameters of the cost function if there are no stock depletion effects on the cost of extraction, or if such effects exist but Z is separable in production from the other inputs. Note that if Z could be included, an estimate of the shadow rental price of one unit of ore in situ, $r\mu - \dot{\mu}$, could be obtained as $-\partial CR/\partial Z$.

$$\begin{aligned}\ln CR &= \alpha_0 + \alpha_Q \ln Q + \sum_i \alpha_i \ln P_i + \alpha_N \ln N + \alpha_T T \\ &\quad + \tfrac{1}{2}\gamma_{QQ}(\ln Q)^2 + \tfrac{1}{2}\sum_i \sum_j \gamma_{ij} \ln P_i \ln P_j + \tfrac{1}{2}\gamma_{NN}(\ln N)^2 \\ &\quad + \tfrac{1}{2}\gamma_{TT}T^2 + \sum_i \gamma_{iQ} \ln P_i \ln Q + \sum_i \gamma_{iN} \ln P_i \ln N \qquad (14) \\ &\quad + \sum_i \gamma_{iT}(\ln P_i)T + \gamma_{QN} \ln Q \ln N + \gamma_{QT}(\ln Q)T \\ &\quad + \gamma_{NT}(\ln N)T, \quad i, j = K, L, E,\end{aligned}$$

where, without loss of generality, $\gamma_{ij} = \gamma_{ji}$. Linear homogeneity in prices is imposed on the cost function by the restrictions[7]

$$\sum_i \alpha_i = 1,$$

$$\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = \sum_i \gamma_{iN} = \sum_i \gamma_{iT} = \sum_i \gamma_{iQ} = 0, \quad i, j = K, L, E.$$

Given the large number of parameters to be estimated, it is desirable to obtain additional effective degrees of freedom by estimating the input cost share equations,

$$M_i = \alpha_i + \sum_j \gamma_{ij} \ln P_j + \gamma_{iT}$$

$$i = K, L$$

jointly with the reproducible cost f[illegible] sum to unity for each observation, o[illegible] deleted from the system of equations to a[illegible] turbance covariance matrix. The remaining equations are estimated using an iterative Zellner-efficient procedure. The parameter estimates are invariant to the choice of equation to be deleted.

Examination of the residuals from the estimated system of equations indicated that the disturbances were heteroscedastic. Therefore we used the heteroscedasticity-consistent covariance matrix estimator proposed by White (1980), $\hat{\mathbf{W}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Omega}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, where $\hat{\mathbf{W}}$ is the estimated covariance matrix, $\mathbf{X}$ is the $n \times k$ matrix of observations on the independent variables, and $\hat{\mathbf{\Omega}} = \text{diag}(\hat{u}_1^2, \ldots, \hat{u}_n^2)$ with $\hat{u}_i$ equal to the residual for the ith observation.

---

[7] Estimation of the model without the restrictions for linear homogeneity in prices imposed indicates that these restrictions result in a significant reduction in goodness of fit. However, since economic theory requires linear homogeneity in prices, the standard practice of imposing it as a maintained hypothesis is followed in subsequent estimation.

Given this estimator for the covariance matrix, linear hypothesis tests of the form $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ are performed using the test statistic $(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'(\mathbf{R}\hat{\mathbf{W}}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})$, where, for tests involving q restricted parameters, $\mathbf{R}$ is a $q \times k$ matrix, $\hat{\boldsymbol{\beta}}$ is the $k \times 1$ vector of parameter estimates, and $\mathbf{r}$ is a $q \times 1$ vector. Under the null hypothesis, the test statistic is asymptotically distributed as χ^2 with q degrees of freedom (White 1980).[8]

Hypotheses to be tested include homotheticity of the production function in reproducible inputs, which is tested by imposing the restrictions $\gamma_{kQ} = 0$, $k = K, L, E$. Homogeneity is tested conditional on the acceptance of homotheticity by imposing the additional restrictions, $\gamma_{QQ} = \gamma_{QN} = \gamma_{QT} = 0$. Hicks's neutral technical change is tested by imposing the restrictions $\gamma_{kT} = 0$, $k = K, L, E, N$.

The reproducible cost function, (14), and cost share equations, (15), are estimated with annual time-series data for the Canadian metal mining industry for 1956 through 1974.[9] Final output, Q, is the dollar value of ore concentrate deflated by the wholesale price index for metal mining. The quantity of capital is calculated using the perpetual inventory method, and the price of capital, P_K, is a modified Christensen and Jorgenson (1969) service price index reflecting acquisition cost, the rate of interest, and the rate of depreciation.[10] The quantity of labor is the total number of workers, and the price of labor, P_L, is equal to average wages plus indirect benefits. Data on energy input are for electricity only. The price of energy, P_E, is equal to expenditure per kilowatt-hour. All input prices are expressed in nominal terms. The quantity of ore extracted, N, is equal to the total number of tons of metallic ore hoisted. All variables entering in log form are normalized such that their values at the means of the data are equal to unity. The time variable, T, is normalized to have the value zero for the median year, 1965.

The results of the hypothesis tests concerning the characteristics of the production function indicate that the production function is both homothetic and homogeneous in reproducible inputs. Hicks's neutral technical change is strongly rejected.[11] We also tested restrictions on several second-order parameters that appeared to reflect multicollinearity among the corresponding variables. The restrictions $\gamma_{TT} = \gamma_{NN} = 0$ could not be rejected either singly or jointly. Since these restrictions increase the precision with which the remaining parameters are estimated, without substantial effects on their estimated values, we have imposed them in the final form of the model.

---

[8] When $\mathbf{r} = 0$ and $\mathbf{R}$ is a $1 \times k$ vector with the ith element equal to unity and the rest zero, the test statistic is the square of a quasi t-statistic for the ith parameter (White 1980).

[9] We are grateful to G. Anders of the Ontario Ministry of Natural Resources for making the data available to us The data are described in more detail in Smithson et al. (1979)

[10] The formula used for the service price of capital is $P_K = q_K(r + \delta)$, where q_K is the acquisition cost of capital, r is the real rate of interest, and δ is the rate of depreciation.

The estimated reproducible cost function is nondecreasing and concave in the prices of reproducible inputs and nonincreasing and convex in the quantity of N, as required for a well-behaved production function (Lau 1976). Of the 20 estimated parameters in the final form of the model, eight are significant at the .01 level.[12] The values of R^2 for the individual equations range from .922 to .995, while the generalized R^2 for the system of equations as a whole is equal to .996.[13]

The estimated shadow price of the resource in situ, $\hat{\mu}$, is calculated for each year as[14]

$$\hat{\mu}_t = -\frac{\partial CR_t}{\partial N_t} = -\Big(\alpha_N + \gamma_{NN} \ln N_t + \sum_i \gamma_{iN} \ln P_{it} + \gamma_{QN} \ln Q_t + \gamma_{NT} T\Big)\frac{\widehat{CR}_t}{N_t}, \quad i = K, L, E. \tag{16}$$

The table shows the corresponding scarcity index normalized to have the value of 100.0 in the 1956 base year. The index indicates a downward trend in natural resource scarcit
son, an index of the real price of fin
ing the metal mining output pric
price index. As shown in the table,
index of natural resource scarcity is
tendency to decrease over time.

---

[11] Anders et al. (1980) use the same basic data to investigate the characteristics of production in this industry. Their specification differs from ours in that a total, rather than reproducible, translog cost function is used. Also, they use a different definition of the technology variable that seeks to reflect the effects of embodied technological change. Despite these differences in specification, their results agree with ours in indicating that the production function is characterized by biased technical change.

[12] When the estimated covariance matrix is not adjusted for heteroscedasticity, 12 of the estimated parameters appear to be significant at the .01 level.

[13] The generalized R^2 is calculated as $1 - (|r_1|/|r_2|)$, where $|r_1|$ is the determinant of the residual moment matrix for the system of eqq. (14) and (15) and $|r_2|$ is this determinant when the coefficients of all slope variables are constrained to equal zero.

[14] In calculating $\hat{\mu}_t$, the predicted values of CR_t implied by the estimated reproducible cost function, eq. (14), are used. Since exponentiation of $(\ln \widehat{CR}_t)$ results in biased estimates of $\widehat{CR}_t$, the transformation suggested by Goldberger (1968, p. 469) is used in calculating $\widehat{CR}_t$.

TABLE 1

NATURAL RESOURCE SCARCITY INDEXES

	Estimated Real Price of Unextracted Ore	Real Price of Final Output*
1956	100.0	100.0
1957	93.5	95.8
1958	103.2	93.0
1959	83.8	94.7
1960	81.4	95.4
1961	81.1	95.5
1962	68.2	94.6
1963	59.9	93.7
1964	53.6	96.1
1965	45.7	99.1
1966	48.3	99.3
1967	44.1	101.4
1968	39.6	102.3
1969	38.1	102.5
1970	33.9	108.8
1971	29.2	105.0
1972	27.5	99.6
1973	19.9	94.7
1974	17.6	98.1

* Wholesale price index for metal mining divided by the general wholesale price index

To obtain a summary description of the differences in the two indexes, we estimated the linear trend equations

$$X_{\mu} = \underset{(16.38)}{103.98} - \underset{(2.93)}{4.77T}, \quad R^2 = .96, \tag{17}$$

$$X_P = \underset{(1.78)}{94.56} + \underset{(0.16)}{0.38T}, \quad R^2 = .26, \tag{18}$$

where X_{μ} is the index of the shadow price of unextracted ore, X_P is the index of the real price of final output, T is time, and the figures in parentheses are estimated standard errors.[15] The trend coefficient is significant at the .10 level in both equations but differs in sign, with the absolute value of the negative coefficient in equation (17) much larger than the positive coefficient in equation (18).

Thus the estimated values of the shadow price of the resource in situ indicate a downward trend in natural resource scarcity, while the price of output indicates a slight upward trend. Which conclusion is "correct" depends on which concept of natural resource scarcity is considered the most relevant in a particular context. However, these results clearly demonstrate the hazards involved in using output price as a proxy for the price of the unextracted resource when the latter measure is the one desired, and therefore the value of having available a procedure for estimating the price of unextracted ore from data for a vertically integrated natural resource industry.

---

[15] The estimated standard errors for eq. (17) are adjusted to account for the fact that $\hat{\mu}$ is an estimate calculated using eq. (16). Without this adjustment, the estimated standard error of the trend coefficient would be only .23

References

Anders, G.; Gramm, W. P ; Maurice, S. C ; and Smithson, C. W. *The Economics of Mineral Extraction* New York: Praeger, 1980.

Barnett, Harold J., and Morse, Chandler *Scarcity and Growth: The Economics of Natural Resource Availability* Baltimore: Johns Hopkins Univ. Press (for Resources for the Future), 1963.

Brown, Gardner M., Jr., and Field, Barry C. "Implications of Alternative Measures of Natural Resource Scarcity." *J P.E* 86, no. 2, pt 1 (April 1978): 229–43.

Christensen, Laurits R., and Jorgenson, Dale W. "The Measurement of U.S. Real Capital Input, 1929–1967." *Rev Income and Wealth* 15 (December 1969): 293–320.

Devarajan, Shantayanan, and Fisher, Anthony C. "Exploration and Scarcity" *J.P E.* 90 (December 1982): 1279–90.

Fisher, Anthony C *Resource and Environmental Economics* Cambridge Cambridge Univ. Press, 1981.

Goldberger, Arthur S. "The Interpretation and Estimation of Cobb-Douglas Functions." *Econometrica* 36 (July–October 1968) 164–72

Halvorsen, Robert, and Smith, Tim R. "On Measuring Natural Resource Scarcity." Mimeographed. Seattle: Univ. Washington, 1984

Hanson, Donald A. "Increasing Extraction Costs and Resource Prices: Some Further Results." *Bell J. Econ.* 11 (Spri

Hotelling, Harold. "Edgeworth's Tax
mand and Supply Functions." *J.P E.*

Johnson, Manuel H., Bell, Frederick W
Resource Scarcity: Empirical Evidence
Econ. and Management 7 (September

Lau, Lawrence J. "A Characterization
Function." *J. Econ Theory* 12 (Februa

Manthy, Robert S. *Natural Resource Commod*
Output, Consumption, Foreign Trade, and Employment [illegible]
1973. Baltimore: Johns Hopkins Univ Press (for Resources for the Future), 1978

Moroney, John R., and Trapani, John M. "Factor Demand and Substitution in Mineral-intensive Industries." *Bell J Econ* 12 (Spring 1981) 272–84.

Pindyck, Robert S. "The Optimal Exploration and Production of Nonrenewable Resources." *J P.E.* 86 (October 1978) 841–61

Silberberg, Eugene. *The Structure of Economics. A Mathematical Analysis* New York: McGraw-Hill, 1978.

Slade, Margaret E. "Trends in Natural-Resource Commodity Prices An Analysis of the Time Domain." *J Environmental Econ and Management* 9 (June 1982): 122–37.

Smith, V. Kerry. "Measuring Natural Resource Scarcity: Theory and Practice." *J. Environmental Econ. and Management* 5 (June 1978): 150–71.

———. "Natural Resource Scarcity: A Statistical Analysis." *Rev. Econ. and Statis.* 61 (August 1979): 423–27. (*a*)

———, ed. *Scarcity and Growth Reconsidered.* Baltimore: Johns Hopkins Univ. Press (for Resources for the Future), 1979. (*b*)

———. "The Evaluation of Natural Resource Adequacy: Elusive Quest or Frontier of Economic Analysis?" *Land Econ.* 56 (August 1980): 257–98.

Smithson, C. W.; Anders, G.; Gramm, W. P.; and Maurice, S. C. *Factor Substitution and Biased Technical Change in the Canadian Mining Industry.* Toronto: Ontario Ministry of Natural Resources, 1979.

White, Halbert. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (May 1980): 817–38.

New Products and the Factor Content of International Trade

Richard A. Brecher and Ehsan U. Choudhri

Carleton University

This paper extends the multicommodity two-input version of the Heckscher-Ohlin model to include "new products," on which considerable interest has been focused by the product-cycle hypothesis and related analyses. The extended model is then used to reexamine the implications of relative factor abundance for the capital and labor content of international commodity trade. As this analysis shows, the presence of new products might in principle resolve the Leontief paradox under certain specific conditions.

I. Introduction

Vernon's (1966) product-cycle hypot
Kravis 1956; Posner 1961; Hufbau
considerable interest in the role of
the factor content of international co
standing the Leontief paradox. Whil
emphasizing the empirical importanc
to be incorporated formally into the stan
world trade, in the tradition of Melvin (196

To address this need, the present paper extends the multicommodity two-input version of the Heckscher-Ohlin model to include new products, defined here as goods whose technology is available in only one country. In Section II, the extended model is used to reexamine

---

The present paper is a condensed version of our (1982c) earlier manuscript, which provides additional details. We gratefully acknowledge helpful comments and suggestions from Jagdish N. Bhagwati, Alan V. Deardorff, Robert C. Feenstra, and Jacob A. Frenkel.

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0001$01.50

the implications of relative factor abundance for the capital and labor content of world trade. As explained in Section III, the presence of new products might in principle resolve the Leontief paradox under certain specific conditions.

II. The Extended Model

Start with the standard Heckscher-Ohlin model of a two-country, two-factor world in which capital and labor are used to produce any number of "old products," defined here as goods whose technology is identical internationally. Extend this model to incorporate one or more new products (defined above). As usual, assume constant returns to scale in production, absence of factor-intensity reversals, identical homothetic tastes for the home and foreign countries, international immobility of capital and labor, perfect competition, and free trade.

To highlight certain stylized features of Vernon's (1966) discussion, we restrict our attention to the interesting case in which the following three assumptions also hold. First, the (physically) capital-abundant country is the sole producer of all new goods.[1] Second, this country has a higher wage-rental ratio than the rest of the world.[2] Third, new products are labor intensive in relation to old goods, when the two types of commodities are compared at the same wage-rental ratio.[3]

Letting the home country be the relatively capital-abundant one,

$$\frac{\bar{K}}{\bar{K}^w} > \frac{\bar{L}}{\bar{L}^w}, \tag{1}$$

where $\bar{K}$ and $\bar{L}$ represent the fixed endowments of capital and labor, respectively, for this economy and $\bar{K}^w$ and $\bar{L}^w$ denote the corresponding endowments for the world as a whole. As we shall now proceed to show, predictions about the factor content of international trade in the extended (unlike the standard) model depend on whether any (old) good is produced in common by both countries.[4]

To dispense first with the simpler case, begin by supposing that there is no "common product," defined here as a good produced in both countries. In this case, the factor requirements per dollar's worth of each good are unique in world production and hence consumption. Therefore, the presence of new commodities does nothing to upset the following condition used by Leamer (1980):

---

[1] See also Krugman (1979) for a (one-factor) model of trade between an innovating "North" and an imitating "South."

[2] If factor prices were assumed to remain equal between countries, introducing new products would not essentially affect the standard factor-content theorems of international trade, as explained by Vanek and Bertrand (1971). See also n. 3 below.

[3] Thus, as could readily be shown, the capital-abundant country cannot have a lower wage-rental ratio than the rest of the world.

[4] On the lack of such dependence when new products are absent, see Brecher and Choudhri (1982*a*).

$$F_c = \alpha \bar{F}^w, \quad F = K, L, \tag{2}$$

where K_c and L_c are the total amounts of capital and labor, respectively, embodied in home consumption and α is the home country's relative share of total world expenditure.

From (1) and (2), $L_c/\bar{L} > K_c/\bar{K}$ or, equivalently, $K_T/\bar{K} > L_T/\bar{L}$, where K_T ($\equiv \bar{K} - K_c$) and L_T ($\equiv \bar{L} - L_c$) denote net exports of capital and labor, respectively, embodied in the commodity trade of the home country. Thus, despite our introduction of new products, there is not yet any violation of Vanek's (1968) well-known proposition concerning the factor-content "chain of comparative advantage," which may be expressed generally as follows: Ranking the factors of a country (say the home one) by relative domestic abundance ($\bar{F}/\bar{F}^w$) is the same as ranking them by relative net exports ($F_T/\bar{F}$).[5]

The situation changes dramatically, however, as we now introduce a common good,[6] produced with factor requirements that differ between countries (given the international divergence in wage-rental ratios). To preserve (2) in the face of these differences, we would clearly need to have equality of θ and θ*, which respectively denote the proportions of home and foreign consumption of the common good met from home production. Given the conventional assumption that international trade of each (ho tional, however, it is easy to see tha home consumption of the common of home production, which (per dol capital and less labor than foreign rental ratio is higher at home than a replaced by

$$F_c = \beta^T \alpha \bar{F}^w, \quad F = K, L, \tag{3}$$

---

[5] More could be said if balanced commodity trade were assumed. Then, letting P_K and P_L denote home (unequal to foreign) prices of capital and labor, respectively, we could readily show that $P_K K_T + P_L L_T < 0$ given cost minimization. From this result and the chain proposition stated above, $L_T < 0$. We could simultaneously have $K_T < 0$, intuitively because the home country's exports of new products embody a "technology content," which might be exchanging internationally for both capital and labor content.

[6] Only one such good is possible in the two-factor, free-trade model with international factor-price differences, as noted by Deardorff (1979). The present analysis extends readily, however, to any number of common products in the presence of production tax-cum-subsidies in favor of import-competing goods.

[7] As the home country is an exporter, nontrader, or importer of the common product, we have $\theta = 1 > \theta^* > 0$, $\theta = 1 > \theta^* = 0$, or $1 > \theta > 0 = \theta^*$, respectively.

where $\beta^K > 1 > \beta^L$ necessarily.[8] If β^K is sufficiently greater than β^L, then from (1) and (2′) we now can have $L_T/\bar{L} > K_T/\bar{K}$, in clear violation of the chain proposition stated above.

The possibility of this violation is illustrated dramatically (with $L_T > 0 > K_T$) in figure 1 for the case of balanced commodity trade. In this diagram, *HH′* and *FF′* are unit isocost lines in the home and foreign countries, respectively. The unit-value isoquants 1, 2, and 3 represent a dollar's worth of output for old goods 1 and 2 and new good 3. (For now, no other commodities exist, and hence isoquant 4 should be ignored.) The slopes of (undrawn) rays *OE* and (flatter) *OD* give, respectively, the capital-labor endowment ratios of the home and foreign countries. For simplicity, (common) good 1 is not traded internationally.[9] Since point *N* lies southeast of point *B*, every dollar's worth of home exports (of good 3) uses more labor and less capital than every equal-valued amount of foreign exports (of good 2). Thus, the home country is a net exporter of labor and importer of capital.

It is important to stress that the relative labor intensity of new products is not itself enough to ensure this violation of the chain proposition. *Given their presence, the crucial issue is the magnitude* of biases (β^K and β^L) created by common-good production. Only if these biases are sufficiently great can the chain be reversed as in the example above.

III. The Leontief Paradox

Although Leamer (1980) has shed considerable light on the Leontief paradox, Leontief's (1954) data for the United States include some features that are not satisfactorily explained in the model with only old products. For instance, as Brecher and Choudhri (1982*b*) have shown, it was paradoxical that $L_T > 0$ (despite $\bar{L}/\bar{L}^w < \alpha$) for the United States. Also, since Leontief's estimates show U.S. capital (labor) requirements per million dollars to be less (greater) for exports than imports, his figures suggest the paradox that $L_T > 0 > K_T$ if commodity trade were balanced.[10]

---

[8] As could readily be shown, $\beta^F - 1 = [(\theta - \theta^*)(a_F - a_F^*)C^*]/\bar{F}^w$, $F = K, L$, where a_K and a_L are the inputs of capital and labor, respectively, per dollar's worth of home output of the common good; a_K^* and a_L^* are the corresponding requirements in the foreign country; and C^* is foreign consumption of the common good. If additional common goods were admitted in accordance with n. 6, we would still have $\beta^K > 1 > \beta^L$.

[9] This simplification was suggested to us by Alan V. Deardorff, who noted that $AE/AN = A'D/A'B$ in this case, since the relative share of (nontraded) good 1 in national income must be the same for both countries (given balanced commodity trade and homothetic tastes)

[10] Admittedly, a move to balanced commodity trade (from the position of surplus in Leontief's data) could have changed the commodity composition of U.S. trade and thus might have altered the factor content per million dollars of exports and imports. There

As is clear from Section II above, an explanation of these paradoxical findings is possible if new products are incorporated into the model. In this case, the labor bias (β^l) due to common-good production might be sufficiently pronounced to yield $\bar{L} > \beta^l \alpha \bar{L}^w$ for the United States, implying that $L_T > 0$ (in view of [2′]). Also, the capital bias (β^K) for the United States might at the same time be great enough to cause $K_T < 0$ under balanced commodity trade (as illustrated above in fig. 1). Thus, *the presence of new products can in principle explain an apparent failure of the Heckscher-Ohlin theorem* in its Melvin (1968)–Vanek (1968) factor-content version, which predicts that $K_T > 0 > L_T$ if commodity trade is balanced.

Finally, with (but not without) new products we can intriguingly have the following new possibility: The Leontief paradox might simply be a measurement error, caused by Leontief's (1954) method of using home rather than foreign requirements to estimate the factor content of U.S. imports.[11] Even when $K_T > 0 > L_T$, we may observe $\hat{L}_T > 0 > \hat{K}_T$, where a circumflex over a variable indicates the Leontief estimate (based wholly on home requirements of factors). Although this combination of true and estimated outcomes cannot arise when all goods are old, the following example illustrates the possibility of this combination in the presence of new products.

In figure 1, assume that good 4 (with unit-value isoquant 4) and good 1 are the world's only commodities. (Thus, isoquants 2 and 3 should now be ignored.) The home country produces and exports (new) product 4, while producing and importing good 1. (Note that
OA′ [not drawn] is now the foreign endowme
lower capital-labor ratio than the dom
still *OE* [not drawn].) Since point *C*
anced commodity trade would imply
however, $\hat{L}_T > 0 > \hat{K}_T$, because poi
(which provides the Leontief estimates
by point *A′*).

As suggested by this example, the hom
exporting capital and importing labor even whe
of factor trade actually holds. For this possible outcome to be consis-

---

is no reason to expect, however, that such a change would have reversed the signs of K_T and L_T. In fact, Baldwin (1971) has estimated the factor content of U.S. trade for a period in which this country's trade surplus was considerably smaller than in Leontief's period. Baldwin's estimates confirm that, per million dollars, the United States was a net exporter of labor and importer of capital.

[11] On the inappropriateness of this estimation procedure when factor prices differ internationally, see Deardorff (1982). Nevertheless, without new products, Brecher and Choudhri's (1982*a*) proof that $K_T > 0 > L_T$ is qualitatively unaffected if factor requirements of home importables are estimated (a la Leontief) by the points where isoquants for such goods touch the unit isocost line of the home country

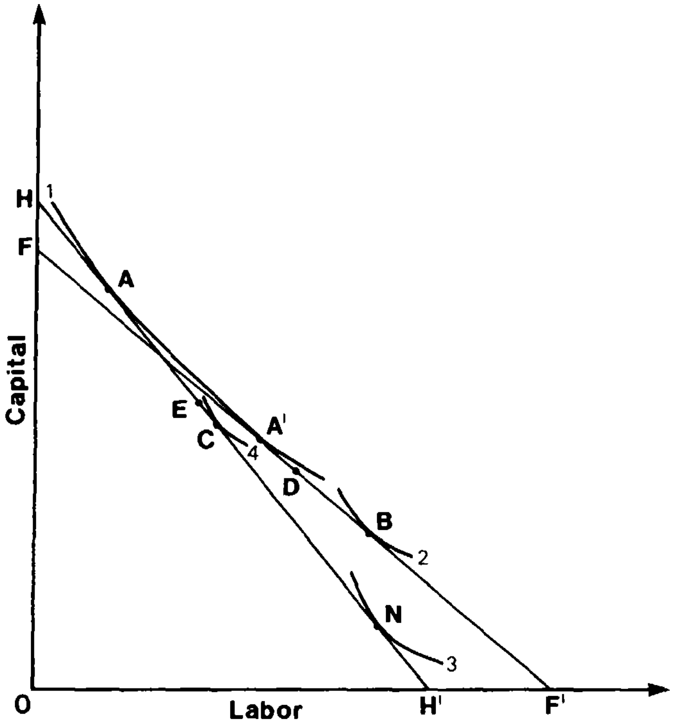


FIG. 1

tent with Leontief's (1954) data, his estimates of labor and capital content per million dollars would have to be only 7 percent below and 20 percent above the true levels, respectively. Although detailed verification of such a possibility would be beyond the scope of the present paper, some readily available data (for labor) suggest that the actual bias may well be considerably greater.[12]

References

Baldwin, Robert E. "Determinants of the Commodity Structure of U.S. Trade." *A.E.R.* 61 (March 1971): 126–46.

Brecher, Richard A., and Choudhri, Ehsan U. "The Factor Content of International Trade without Factor-Price Equalization." *J. Internat. Econ.* 12 (May 1982): 277–83. (*a*)

---

[12] For evidence on differences in (direct) labor coefficients between the United States and three other countries (India, Japan, and the United Kingdom), see Lary's (1968, app. B) detailed bilateral comparisons of value added per employee.

———. "The Leontief Paradox, Continued." *J.P.E.* 90 (August 1982): 820–23. (*b*)

———. "New Products and the Factor Content of International Trade." Carleton Economic Papers no. 82-20. Ottawa: Carleton Univ., November 1982. (*c*)

Deardorff, Alan V. "Weak Links in the Chain of Comparative Advantage." *J. Internat. Econ.* 9 (May 1979): 197–209.

———. "The General Validity of the Heckscher-Ohlin Theorem." *A.E.R.* 72 (September 1982): 683–94.

Hufbauer, Gary C. *Synthetic Materials and the Theory of International Trade.* Cambridge, Mass.: Harvard Univ. Press, 1966.

———. "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods." In *The Technology Factor in International Trade*, edited by Raymond Vernon. New York: Columbia Univ. Press (for N.B.E.R.), 1970.

Kravis, Irving B. "'Availability' and Other Influences on the Commodity Composition of Trade." *J.P.E.* 64 (April 1956): 143–55.

Krugman, Paul. "A Model of Innovation, Technology Transfer, and the World Distribution of Income." *J.P.E.* 87 (April 1979): 253–66.

Lary, Hal B. *Imports of Manufactures from Less Developed Countries.* New York: Columbia Univ. Press (for N.B.E.R.), 1968.

Leamer, Edward E. "The Leontief Paradox, Reconsidered." *J.P.E.* 88 (June 1980): 495–503.

Leontief, Wassily W. "Domestic Production and Foreign Trade: The American Capital Position Re-Examined." *Econ. Internaz.* 7 (February 1954): 9–38. (First appeared in *Proc. American Philosophical Soc.* 97 [September 1953]: 332–49.)

Melvin, James R. "Production and Trade with Two Factors and Three Goods." *A.E.R.* 58 (December 1968): 1249–68.

Posner, Michael V. "International Trade and Technical Change." *Oxford Econ. Papers* 13 (October 1961): 323–41.

Vanek, Jaroslav. "The Factor Proportions Theory: The N-Factor Case." *Kyklos* 21, no. 4 (1968): 749–54.

Vanek, Jaroslav, and Bertrand, Trent J.
Commodity World." In *Trade, Balan*
International Economics in Honor of Char
N. Bhagwati et al. Amsterdam: North-

Vernon, Raymond. "International Invest
Product Cycle." *Q.J.E.* 80 (May 1966)

Angell and the Stable Money Rule

Joong-Koon Lee and Donald C. Wellington
University of Cincinnati

The stable money rule is part of monetarism. That aspect of the rule may account for monetarists' being chary of acknowledging the wide spectrum of advocacy of the rule. In particular, the rule's early advocacy by James W. Angell has not been given sufficient recognition.

Two decades ago a number of writers, Howard S. Ellis (1964, p. 202), Harry G. Johnson (1965, p. 31), and Richard Selden (1966, p. 328), took special note of a later article (Angell 1960) and uniformly echoed the view that Angell had belatedly joined the stable money school. The same view seems to have been held, at least implicitly, by Thomas Humphrey. In his study of the non-Chicago economists' contributions to the American quantity theory tradition, Humphrey (1971) found no occasion to mention Angell's name throughout his entire study. Later, in response to Patinkin's (1973, p. 456) strongly worded criticism of his failure to recognize Angell's contribution, he conceded: "As Patinkin rightly points out, James Angell . . . pioneered in the . . . uses of the income velocity concept and thus deserves recognition as one of the important non-Chicago quantity theorists writing in the 30's" (Humphrey 1973, p. 461). While conceding on this issue of income velocity, Humphrey continued to ignore Angell's long-standing support of the stable money proposition, thereby creating the impression that Ellis et al. were essentially correct in their view that Angell joined the stable money school only in 1960.

The purpose of this note is to show that Angell had been a proponent of a stable money rule ever since the early 1930s. In order to establish this point, we call attention in Section I to some of Angell's earlier writings. In Section II, we comment on some related matters.

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0010$01.50

I

In December 1932, Angell delivered a paper before the American Association for the Advancement of Science (Angell 1933*a*), in which he adumbrated his views on "Monetary Prerequisites for Employment Stabilization."[1] These views, later incorporated into another essay (Angell 1933*b*), were quite specific on the stable money issue. He argued that "the total net quantity of cash and bank deposits" should be held "as nearly constant as possible" in the short run. Beyond the short run, however, "the constancy of the net quantity of cash and bank deposits should be modified for estimated enduring changes in the circular velocity of money itself, so that the product of money stock times circular velocity remains substantially constant . . . [and] should . . . be allowed to vary with population, to prevent substantial changes in average money income per caput" (Angell 1933*b*, p. 58).

Angell thought that these criteria would give the central bank authorities "a simple and . . . a reasonably workable guide for their actions" (Angell 1933*b*, p. 59). In implementing such a policy, however, the authorities should be aware of the fact that the principal source of variation in the money supply (or "the supply of media of exchange," as he called it) came from bank deposits. In order to keep bank deposits at the desired level, "it is principally necessary only to control the aggregate size of Federal Reserve member banks' reserve balances" by employing discount and open-market policies (Angell 1933*b*, pp. 65–66).

Soon afterward came one of the most memorable years in the annals of the economics profession. It was 1935, the year when the foremost living quantity theorist, Irving F
published the book *100% Money*, a
tained the proposition that the qu

---

[1] As far as we know, this work is the earli
money growth. It came fully 4 years before
year before the celebrated mimeographed
Reform" that was being circulated among Chicago
Simons 1936, p. 4). Although the stable money rule
ancient heritage, it is perhaps best to view it as a twentieth-century proposal in that its current form probably stems from the ending of the gold standard and the coming of very active central banking. Previous forms early in the nineteenth century (Tavlas 1977) are more descriptions of a desirable state than meaningful proposals for central bank action and differ from yet earlier ideas in being set in a growing instead of a stationary economy. Implications of that contrast are merely that the rule be a constant rate of growth rather than a constant level. Ignoring such a minor modification suggests that the true intellectual origins of the stable money rule go back at least to Suetonius (A.D. 70–122), who remarked on the monetary consequences of the Emperor Augustus's triumphs, "Nam et invecta urbi Alexandrino triumpho regia gaza tantam copiam nummariae rei effecit, ut faenore deminuto plurimum agrorum pretiis accesserit" (Suetonius, Divus Augustus 41.1).

and contract as business expands and contracts" with some allowance being made for the expansion of the money stock over time (Fisher 1935, pp. 160 and passim). Angell's reaction was immediate: "This is the doctrine, of the 'elasticity' of currency and credit, which is at the heart of the Federal Reserve plan . . . [and] which I think should be expunged once for all from the economic organon" (Angell 1935, p. 25). He repeated his call for a stable money growth linked only to secular changes in circular velocity and population (Angell 1935, pp. 26, 34). According to him, these two trend factors would have required "an increase in the stock of money of less than 1.5 per cent a year" (Angell 1935, p. 34) during the decade of 1920–30.

Angell proposed his own "100 percent reserve plan," which he believed would achieve the desired results while remaining free of the defects of other such plans. His plans required "the conversion of the [existing] demand deposit liabilities of the commercial banks into similar liabilities of the United States, but backed 100 per cent by currency and administered by the commercial banks as agents" (Angell 1935, p. 34). Such a system, he thought, would "permit a rational control of the total size of the stock of money itself" free from transitional complications and subsequent political abuses (Angell 1935, p. 35). Angell, however, did not consider the establishment of a 100 percent reserve system important in itself. Its virtue was that it served as a means to "secure reasonable behavior of the money stock."

Angell's *Behavior of Money* was published in 1936, the same year that witnessed the publication of Henry Simons's "Rules versus Authorities in Monetary Policy." Angell's book, which Howard Ellis called "a work of commanding significance in monetary science," summarized the results of a statistical analysis involving money, national income, and velocity that had been carried out over "several years." From his analysis (Angell 1936, pp. 161–62) that showed that the quantity of money (especially that of deposits) had no close correlation with the short-run pace of real economic activity, he concluded that any attempt to achieve "desirable general economic objectives by enforcing artificial and . . . frequent *fluctuations* in the quantity of money" would encounter "serious difficulties" (Angell 1936, pp. 161–62). In contrast, he found that money income displayed a close association with "[changes and trends] of the stock of circulating money" in 1909–28, with the circular velocity varying within a narrow range around a horizontal trend (Angell 1936, p. 162). These results led Angell to state: "if the ultimate objective of policy is to induce a greater degree of stability in national and individual money incomes . . . then the most effective procedure is to stabilize the quantity of money. This quantity would not be held absolutely constant . . . but

. . . allowed to change . . . gradually and evenly" (Angell 1936, p. 163).[2]

Unlike quite a number of other economists, Angell never thought that the stable money rule could be a cure-all. In particular, he warned, "marked fluctuations in economic activity and in the national income would doubtless still remain, but there is ground for thinking that they would be far less severe and injurious than those from which we now suffer" (Angell 1936, p. 163).

Angell's most comprehensive discussion of stable money growth came in an essay written in honor of Irving Fisher. He pointed out that economists generally agreed that monetary policy should be so planned and operated as to make the economy cope more efficiently with "large and small booms and collapse" and to ensure a reasonably full utilization of resources at all times. But opinions differed "on just what the specific content of that policy should be." For Angell the choice was clear. "If we take national money income as a rough guide to the total money volume of general economic activity, it therefore seems plausible to think that an easy method for getting at the larger fluctuations in this total activity, and one not hitherto tested deliberately on any extensive scale, can perhaps be developed by controlling the quantity of circulating money itself" (Angell 1937, p. 55). Then he referred to the stable money plan that he had proposed in earlier writings, stabilizing the stock of money and allowing it to increase only with population growth and long-term velocity changes. Finally he discussed the institutional arrangements appropriate to ensure the controlled growth of the money stock. He urged the adoption of his
100 percent reserve plan, recognizing t
the existing banking practice, but "
practical terms to the attempt to
through such more familiar devic
open-market operations and changin
1937, p. 56, n. 5). Such devices re
powers" against many odds including
[that] comes in slowly" (Angell 1937, p. 56,
percent reserve plan would eliminate "nearly all discretion" and could
be "semiautomatic and positive in its action" (Angell 1937, p. 56, n. 5).

---

[2] Angell's proposal was not well received by Lauchlin Currie of Harvard who, in his review of this book, said: "Whether this [the stabilization of per capita money income] could be accomplished by stabilizing the volume of money, as Professor Angell implies, is a highly debatable question" (Currie 1936, p. 791). Frederick A. Bradford's review also recalled Angell's earlier criticism of the "doctrine of the elasticity of currency and credit" and expressed his strong objection to taking "the problem of monetary control as almost exclusively quantitative problem" (Bradford 1937, p. 371)

Although Angell argued very strongly in favor of stable monetary growth,[3] he did not have any illusion that such growth by itself would "end all fluctuations in the total money volume of economic activity." But "if the quantity of money were kept reasonably stable," Angell concluded, "these remaining and perhaps irreducible quasi cycles would become relatively small in size, and would hence be of secondary importance" (Angell 1937, p. 83).

II

It is clear from the preceding section that Angell began advocating a stable money rule from the early 1930s. His early advocacy was duly recognized by Joseph E. Reeve, who, in a survey of monetary reform movements in the United States, credited Angell and Henry Simons for making "the most important of the relatively few suggestions for stabilization of the quantity of money" (Reeve 1943, p. 330). Angell's call for stable money growth in 1960 was not a new step but one that he had consistently and continuously maintained since the 1930s.[4]

As has been pointed out, Humphrey erred in totally overlooking Angell's contribution to the stable money rule, a cause that became the most celebrated of Chicago school causes. Humphrey needed a non-Chicago-school economist for his case. Instead of Angell, he chose Lionel Edie and Carl Snyder because of their "fixed growth rate rule." Both Edie and Snyder wrote extensively on such a rule. What they were proposing, however, was not a fixed *money* growth rate rule but a fixed *credit* growth rate rule. Nowhere in the sources that Humphrey cites did either Edie or Snyder make specific references to money supply (or to currency plus net deposits or some such magnitudes) as such; their references were invariably to credit or to credit-related magnitudes. This emphasis, we believe, was also why Reeve (1943, p. 330) classified Edie and Snyder as advocates of the proposal to stabilize "the ratio between the volume of credit and the secular expansion of industry and trade," which is distinct from either the proposal to stabilize the price level or the proposal to stabilize the quantity of money. Edie and Snyder can only be considered proponents of the stable money rule if there always exists a one-to-one relationship between bank credit and the money supply and the former can be considered a proxy for the latter. In fact, bank credit is not perfectly correlated with the money supply and the proposals to stabilize them are not identical.[5]

---

[3] Angell again dwelt briefly on the need for controlled growth of money in his book (1941, p 256) Shortly afterward, he was joined by Warburton, who, besides sharing a monetarist approach, expressed the same views on stable money growth. Unlike Angell, he did not so much make policy statements of these views as incorporate them in his analysis (Warburton 1945, p 121; 1946, pp. 442–44).

[4] The only noteworthy change in the 1960 article is that Angell accepted the idea of switching money from being linked to population growth and secular changes in velocity to being linked to the "real growth rate of the economy" and by proposing a number of exceptions to the rule. Those exceptions included such measures as to meet seasonal influences or needs arising out of large Treasury financing, to counteract "disorderly conditions" developing in the government securities market, or to counter the threat of a money market panic (Angell 1960, pp. 251–52).

Why Angell's early advocacy of the stable money rule has been so completely overlooked in the current literature is an intriguing question. He wrote about it, not once or twice casually, but repeatedly over an extended period of time in major journals, monographs, and contributed essays.[6] We respectfully submit that monetarists should, albeit belatedly, receive James W. Angell into the fold.

References

Angell, James W. "Monetary Prerequisites for Employment Stabilization." In *Stabilization of Employment*, edited by Charles F. Roos. Bloomington, Ind Principia, 1933. (*a*)

———. "Monetary Control and General Business Stabilization." In *Economic Essays in Honour of Gustav Cassel*. London: Allen & Unwin, 1933. (*b*)

———. "The 100 Per Cent Reserve Plan." *Q J E.* 50 (November 1935): 1–35.

———. *The Behavior of Money: Exploratory Studies*. New York: McGraw-Hill, 1936.

———. "The General Objectives of Monetary Policy." In *The Lessons of Monetary Experience. Essays in Honor of Irving Fisher*, edited by Arthur D. Gayer. New York: Farrar & Rinehart, 1937.

———. *Investment and Business Cycles*. New York. McGraw-Hill, 1941.

———. "Appropriate Monetary Policies and Operations in the United States Today." *Rev. Econ. and Statis.* 42 (August 1960): 247–52.

Bradford, Frederick A. "Angell, the Behavior of M[illegible] 1937): 364–73.

---

[5] This issue is not a dated one. Although
more of a policy rule in the sense that at least
used as intermediate targets, economists have
measure of credit liabilities that includes bank
general economic activity as well as being
Reserve System. The choice need not be between
could conceivably be used in conjunction (Friedman
significance, bank credit, as distinct from hat money, perhaps [illegible]
making of transactions and the shift to its use may account for many of the notable features of economic development (Townsend 1983).

[6] His writings were not badly received by famous monetarists. Friedman and Schwartz, in their discussion of how the banking crises should have been handled, have written: "Some academic people, such as Harold Reed, Irving Fisher, J. W. Angell, and Karl Bopp expressed similar views. Despite these important exceptions, the literature and particularly the academic literature, on the banking and liquidity crisis is almost as depressing as that on the contraction in general" (Friedman and Schwartz 1963, pp. 409–10).

Currie, Lauchlin. Review of *The Behavior of Money: Exploratory Studies*, by James W. Angell. *A.E.R.* 26 (December 1936): 789–91.

Ellis, Howard S. "Limitations of Monetary Policy." In *United States Monetary Policy*, rev. ed., edited by Neil H. Jacoby. New York: Praeger, 1964.

Fisher, Irving. *100% Money*. New York: Adelphi, 1935.

Friedman, Benjamin M. "Empirical Issues in Monetary Policy: A Review of Monetary Aggregates and Monetary Policy." *J. Monetary Econ.* 3 (January 1977): 87–101.

———. "The Role of Money and Credit in Macroeconomic Analysis." In *Macroeconomics, Prices, and Quantities: Essays in Memory of Arthur M. Okun*, edited by James Tobin. Washington: Brookings, 1983.

Friedman, Milton, and Schwartz, Anna J. *A Monetary History of the United States, 1867–1960*. Princeton, N.J.: Princeton Univ. Press (for N.B.E.R.), 1963.

Humphrey, Thomas M. "Role of Non-Chicago Economists in the Evolution of the Quantity Theory in America, 1930–1950." *Southern Econ. J.* 38 (July 1971): 12–18.

———. "On the Monetary Economics of Chicagoans and Non-Chicagoans: Reply." *Southern Econ. J.* 39 (January 1973): 460–63.

Johnson, Harry G. "Monetary Theory and Policy." In *Surveys of Economic Theory*. Vol. 1. London: St. Martin's (for American Econ. Assoc.), 1965.

Patinkin, Don. "On the Monetary Economics of Chicagoans and Non-Chicagoans: Comment." *Southern Econ. J.* 39 (January 1973): 454–59.

Reeve, Joseph E. *Monetary Reform Movements: A Survey of Recent Plans and Panaceas*. Washington: American Council on Public Affairs, 1943.

Selden, Richard T. "Stable Monetary Growth." In *Monetary Theory and Policy*, edited by Richard A. Ward. Scranton, Pa.: International Textbook Co., 1966.

Simons, Henry C. "Rules versus Authorities in Monetary Policy." *J.P.E.* 44 (February 1936): 1–30.

Suetonius, Tranquillus C. *De Vita Caesarum*. Edited by Maximilian Ihm. Stuttgart: Teubner, 1958.

Tavlas, George S. "Some Initial Formulations of the Monetary Growth-Rate Rule." *History Polit. Econ.* 9 (Winter 1977): 535–47.

Townsend, Robert M. "Financial Structure and Economic Activity." *A.E.R.* 73 (December 1983): 895–911.

Warburton, Clark. "Monetary Theory, Full Production, and the Great Depression." *Econometrica* 13 (April 1945): 114–28.

———. "Quantity and Frequency of Use of Money in the United States, 1919–45." *J.P.E.* 54 (October 1946): 436–50.

Comment

Aggregation and the Factoral Content of Trade

Richard Gift

University of Kentucky

William Marxsen

BellSouth Corporation

With the publication of Vanek's pioneering extension (1968) of the Heckscher-Ohlin theorem to many goods and factors, it became clear that paradoxical trade patterns were possible in a well-behaved neoclassical world. In a recent article in this *Journal*, Edward Leamer (1980) has employed Vanek's model to demonstrate that Leontief's paradoxical trade patterns are consistent with a very restrictive set of Heckscher-Ohlin-Vanek (HOV) assumptions.[1] His analysis suggests that Leontief's and other investigators' appeals to ad hoc assumptions
to explain the Leontief paradox are
overly restrictive view of the model

In the general HOV model emplo
capital and labor are two of the man
and used in production. But an issu
that would seem to be equally impor
ber of factors in the analysis is the mo
factors are aggregated. Clearly, capital an
factors, rather than each being a single factor of production.

The first version of this paper was written while Marxsen was at the University of Kentucky. We have benefited from numerous comments from David Clifton, an anonymous referee, and the International Economics Workshop at the University of Kentucky. Any remaining errors or confusions are our responsibility.

[1] In a largely overlooked paper, Williams (1970) used a slightly different framework to establish this fact as well.

[*Journal of Political Economy*, 1984, vol. 92, no. 5]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9205-0008$01.50

The purpose of this paper is to extend Leamer's analysis to include aggregate factors. It will be shown that the method of aggregation will greatly affect the HOV predictions.

The Basic Model

All of the assumptions of the factor price equalization theorem hold, and, in addition, there are identical and homothetic tastes worldwide. As a consequence, the derived demand for factors in the home country is a fraction of the world demand, which in general equilibrium equals the world factor supply. Thus the country's net factor exports equal its endowment minus a fraction of the world's factor endowment, or

$$E = F - \lambda F^*, \tag{1}$$

where E, F, and F^* are column n-vectors of net factor exports, the domestic factor endowment, and the world factor endowment, respectively, and λ equals Y/Y^*, or the ratio of domestic national income to total world income (hereafter, all asterisked variables will be world values).

The Heckscher-Ohlin theorem follows directly from condition (1). For any given factor, say i, condition (1) implies[2] that

$$E_i \gtreqless 0 \text{ as } F_i \gtreqless \lambda F_i^* \text{ or } \frac{F_i}{Y} \gtreqless \frac{F_i^*}{Y^*}. \tag{2}$$

Whenever $F_i/Y > F_i^*/Y^*$, factor i is abundant and will be exported (net). Furthermore, the relative trade patterns of any two factors may be determined from condition (1) as well. For any factors i and j, it follows from straightforward manipulation of the ith and jth equality in (1) that

$$\frac{E_i}{F_i} \gtreqless \frac{E_j}{F_j} \text{ as } \frac{F_i}{F_i^*} \gtreqless \frac{F_j}{F_j^*} \text{ or } \frac{F_i}{F_j} \gtreqless \frac{F_i^*}{F_j^*}. \tag{3}$$

Factor i is abundant relative to factor j in the home country whenever the country's endowment of i per unit of j exceeds that for the world, and net exports of i as a fraction of the country's endowment will exceed that fraction for j. Leamer has confirmed condition (3) for capital and labor using Leontief's original (1953) data.

Another way to gauge factor abundance and factoral trade is to use income measures of factors.[3] Let R represent a diagonal matrix whose nonzero elements are the various worldwide factor prices (R_i: $i = 1, \ldots, n$). Both sides of condition (1) can be multiplied by R to yield

---

[2] In a recent paper, Brecher and Choudhri (1982) have shown that Leontief's original data fail to conform to this condition and hence reopen the paradox.

[3] Harkness (1978) developed the income approach in a more general model and employed it to examine more recent data.

$$RE = RF - \lambda RF^*. \tag{4}$$

The income version of the Heckscher-Ohlin theorem can now be shown for each factor and for pairs of factors by solving for conditions analogous to (2) and (3). For notational convenience, define $Y_i = R_iF_i$. Then,

$$R_iE_i \gtreqless 0 \text{ as } Y_i \gtreqless \lambda Y_i^* \text{ or } \frac{Y_i}{Y} \gtreqless \frac{Y_i^*}{Y^*} \tag{5}$$

and

$$\frac{R_iE_i}{Y_i} \gtreqless \frac{R_jE_j}{Y_j} \text{ as } \frac{Y_i}{Y_i^*} \gtreqless \frac{Y_j}{Y_j^*} \text{ or } \frac{Y_i}{Y_j} \gtreqless \frac{Y_i^*}{Y_j^*}. \tag{6}$$

With factor price equalization, conditions (5) and (6) produce exactly the same rankings of factors from most abundant to most scarce as do conditions (2) and (3), and hence the two pairs of conditions may be used interchangeably in empirical tests of the model. The income version's main appeal, however, is its apparent avoidance of many practical problems in the measurement of factor intensities. In practice, reliable data concerning numbers of machines or quantities of other factors are unavailable, and hence measuring capital and other factor services by income flows represen
this manner, "aggregation problem
valuations in weighting qualitatively
But such aggregation violates the
identical and homogeneous and thu
model in predicting trade flows. Jus
aggregation is an important question

The Effects of Aggregation

Let A represent a proper subset of the n distinct factors that contains at least two factors chosen for aggregation. For notational simplicity, assume that A contains the first through the mth individual factors and that there is a linear aggregation operator that assigns some weight s_i, $i = 1, \ldots, m$, to each of the individual factors in A, or

$$A = s_1F_1 + s_2F_2 + \ldots + s_mF_m. \tag{7}$$

Assume that an identical procedure is followed for A^* for the world as a whole. Likewise, employ the same weights for net exports of the aggregate, ϵ,

$$\epsilon = s_1E_1 + s_2E_2 + \ldots + s_mE_m. \tag{8}$$

Both sides of equation (1) for an individual factor can be multiplied by the appropriate weights and the resulting equations for the individual elements in the aggregate factor can be added together to yield an identical condition for the aggregate. It immediately follows that the physical-factor-flow Heckscher-Ohlin conditions (2) and (3) will hold for aggregate as well as distinct factors.

What makes the result above interesting is that no limitations have been imposed on the weights, s_i. Any linear aggregation scheme that is consistently applied to home and world endowments and trade flows will yield Heckscher-Ohlin behavior. For instance, if an investigator wishes to aggregate a number of resources under the heading of land, and the specific factors are such things as farmland, fresh water, and coal reserves, then any weighting scheme will yield Heckscher-Ohlin behavior. If the investigator arbitrarily counts one acre of farmland as 10 units of land, a cubic mile of fresh water as one unit of land, and a million tons of coal reserves as one-tenth unit of land, the resulting trade flow measures will still be consistent with the Heckscher-Ohlin theorem. Indeed, any specific factor could be given a negative weight in an aggregate factor without altering the predictions.

This finding is intuitively troubling, since, for any given factor endowments and specific trade patterns, all the individual E_i's are uniquely determined, while ϵ will change in magnitude and perhaps sign as aggregation weights are arbitrarily changed. That the sign and magnitude of ϵ depend on the s_i values chosen would appear to destroy the predictive power of the model. But changing the s_i terms alters the home country and world endowments of the aggregate factor as well. Any change in weights that changes the sign of ϵ also changes its relative abundance in the home country.

For the aggregate factor, A, there will be an associated "factor price," V, and total factor income, Z, such that the aggregate factor income equals the sum of the individual factor incomes,

$$Z = VA = \Sigma R_iF_i, \tag{9}$$

for all i specific factors in the aggregate. Once aggregation weights (s_i) have been chosen, the value of V will depend on the individual factor prices as well as on the quantities of the individual factors. Since individual factor proportions will differ at home and abroad, equalization of individual factor prices will not imply aggregate factor price equalization. An immediate implication of the unequal aggregate factor prices is that the income and physical versions of the theorem will no longer yield identical rankings of abundance and trade-flow patterns. Thus a country might be labor poor by head count and yet rich in labor power as measured by total wage income.

A straightforward application of these results is to reexamine the Leontief (1953), Leamer (1980), and Brecher-Choudhri (1982) findings under different methods of aggregating labor. In all of their cited papers, the authors aggregate by head count. The analysis presented here has shown that such a procedure is legitimate and should produce trade measures consistent with conditions (2) and (3). Thus, with aggregation, Leamer's finding supports the model, and Brecher and Choudhri's does not.

Still untested with the HOV model is Leontief's contention that the United States was human capital abundant. As the analysis has demonstrated, any arbitrary way of measuring human capital will be consistent with the model.[1] But perhaps the most reasonable approach is to measure labor abundance and trade flows via income and thus equate earnings with human capital. When Leontief's original factoral trade calculations are adjusted to measure labor via earnings rather than head count, the relevant term, ϵ_L/L, rises by approximately 10 percent, implying increased labor abundance. But this is not nearly enough to reverse Leamer's findings that $\epsilon_K/K > \epsilon_L/L$, and thus, contrary to Leontief's claim, the data imply that physical capital was more abundant than human capital.[2]

Context

Once the possibility that observed
aggregation is admitted, it follows
an aggregate's average price may ex
zation. With factor price equaliza
implies a relative abundance of the
nents of the aggregate and not a
factor. Hence, when individual factor prices
tor price disparities will impart no information concerning aggregate
abundance. Thus, aggregation, which has been shown to be harmless
in the present model, is quite perilous when one employs (aggregate)
factor prices themselves to predict trade flows.

This claim may be extended to autarky prices as well. In autarky, individual factor prices will generally be unequal and their relative values will indicate abundance. For any country, those individual factors whose autarky prices exceed world averages will be scarce and imported via commodity trade. However, if any aggregation of different individual factors is involved in the determination of factor prices, those factors prices may be "contaminated" by different individual factor proportions within the aggregate. Thus a high autarky price may be explained either by the scarcity of the aggregate or by different individual factor proportions within the aggregate.

---

[1] Kenen (1965) employed an aggregation procedure that breaks labor into skilled and unskilled components and then adds the capitalized value of the skill to the capital stock. Even this procedure is consistent with the model.

[2] With the human capital approach, the Brecher-Choudhri paradox still remains.

These aggregation effects have profound implications for economic development planning and policies. If planners base growth strategies on average factor prices, their policies may be "against the wind," in the sense of being contrary to, rather than in line with, relevant factor endowments. Likewise, trade policies themselves may be similarly quixotic.

The analysis has attempted to offer further explanation of factor intensity paradoxes while retaining the spirit of the HOV model, which is that factor endowments are the major determinants of the pattern of comparative advantage.

References

Brecher, Richard A., and Choudhri, Ehsan U. "The Leontief Paradox, Continued." *J P E.* 90 (August 1982): 820–23.

Harkness, Jon P. "Factor Abundance and Comparative Advantage." *A E.R.* 68 (December 1978): 784–800

Kenen, Peter B. "Nature, Capital, and Trade." *J.P E.* 73 (October 1965): 437–60.

Leamer, Edward E. "The Leontief Paradox, Reconsidered." *J.P.E.* 88 (June 1980): 495–503.

Leontief, Wassily W. "Domestic Production and Foreign Trade: The American Position Re-Examined." *Proc. American Philosophical Soc* 97 (September 1953): 332–49. In *Readings in International Economics*, edited by Richard E. Caves and Harry G. Johnson. Homewood, Ill.: Irwin, 1968.

Vanek, Jaroslav. "The Factor Proportions Theory: The N-Factor Case." *Kyklos* 21, no. 4 (1968): 749–56.

Williams, James R. "The Resource Content in International Trade." *Canadian J. Econ* 3 (February 1970): 111–22.

Book Reviews

The Economics of Housing Vouchers By JOSEPH FRIEDMAN and DANIEL H. WEINBERG.
New York: Academic Press, 1982. Pp. 218. $27.50.

This book reports the major results of the Demand Experiment component of the Experimental Housing Allowance Program (EHAP) conducted by the U.S. Department of Housing and Urban Development during the 1970s. The purpose of EHAP was to see if direct cash payments to households were an "acceptable" alternative to public housing and similar subsidized construction programs as a means of improving the housing occupied by the poor. In the Demand Experiment, some 1,200 households in each of the Pittsburgh and Phoenix SMSAs were given vouchers to spend for housing in the private market under a variety of subsidy formulas. A large amount of data were collected on the behavior of these households; Friedman and Weinberg, then working at Abt Associates, spent several years estimating demand elasticities. This book brings together much of their work, a large part of which has been available previously only in government contract reports.

Their principal findings are that both the price and income elasticities were very low, about −0.2 and 0.35, respectively. This is based on a sample of households that received an unrestricted price subsidy, ranging from 20 to 60 percent, and a nearly equal number of unsubsidized control households. The data are fitted using a Stone-Geary utility function. The observation period covered the first 2 years after the househ[illegible] mates were about the same for those h[illegible] subsidies and those that stayed in their [illegible] ently made a once-for-all change in t[illegible] equilibrium position rather than gradua[illegible] was not much affected by the allowan[illegible] were not careful shoppers; about one-h[illegible] ditures were "wasted" because they pa[illegible] sample of households, a minimum quality sta[illegible] tion of receiving the subsidy. This led some of th[illegible] ing the standards, but they did not increase overall housing consumption (measured by expenditures), and many other households did not select standard housing and thus received no subsidy. Another sample was offered subsidies for meeting a minimum rent standard, to see if it would serve as a proxy for quality; most subsidized households chose a unit that did not meet the minimum quality standards. The minimum rent standard also induced overpayment by subsidy recipients.

As is apparent from this summary, the scope of the book is narrower than

[*Journal of Political Economy*, 1984, vol. 92, no. 5]
© 1984 by The University of Chicago. All rights reserved

its title implies. It does not discuss the supply response to vouchers, which was studied separately in the Supply Experiment component of EHAP. Nor is it a full report on the Demand Experiment; it omits the interesting and politically important issue of participation, for example. Even as an empirical study of the demand for housing by low-income households, it ignores the physical dimension of housing. This means that it bypasses some fascinating issues. For instance, which housing characteristics were valued by program administrators but not by subsidy recipients? Conversely, which ones were valued by recipients and by the market, but not by administrators? The book will therefore appeal chiefly to economists who are estimating housing demand elasticities. They will find it a useful reference; the technical analysis is carefully done, and a number of methodological issues, such as selection biases in the samples, are pursued in detail. Those wanting a broader evaluation of the housing allowance experiments will be better served by any of three recent overview volumes (Bradbury and Downs 1981; Struyk and Bendick 1981; Friedman and Weinberg 1983).

But while Friedman and Weinberg could have considered more issues, what they have done in this book is itself important and has attracted quite a bit of professional attention. The central tendency of most recent empirical work is around 0.7 or slightly larger in absolute value for both price and income. This figure has been declining gradually—it is well below the consensus of a decade ago—and there is some agreement that elasticities are lower for renters and perhaps for the poor. But even taking these modifications into account, Friedman and Weinberg's estimates are at the low end of the distribution. Their results also appear to be at variance with the historical pattern of housing consumption changes. For example, between 1950 and 1980 real rents fell by 10–15 percent and real renter incomes roughly doubled. Friedman and Weinberg's elasticities imply that per household consumption should have risen by about a quarter; in fact, it rose by two-thirds. The actual changes are consistent with elasticities about double their estimates. A possible explanation is the changing demographic characteristics of renters; they are increasingly single-parent families, individuals living alone, or the elderly. However, Friedman and Weinberg also find that elasticities are about the same for various demographic groups, leaving the inconsistency unresolved.

A number of economists have attempted to explain the differences between these estimates and the bulk of the literature, without much success. Unfortunately, Friedman and Weinberg themselves do not address this subject. They merely review the literature, emphasizing the lowest estimates in previous work, before presenting their own findings. They thus minimize the differences without attempting to reconcile them.

One issue that has attracted attention is the time path of adjustment in response to a change in permanent income. The Friedman-Weinberg analysis uses the first 2 years of the experimental data, whereas the limited empirical literature suggests that households take several years longer to adjust fully. Thus it may be that Friedman and Weinberg are observing only the first part of an adjustment process, in which case it is very hard to estimate the long-run response.

Friedman and Weinberg argue that the pecuniary and psychic costs of moving are so high that households try to make just one move to achieve their new equilibrium position rather than reaching it gradually over time. It is also true that most stock adjustment models have used aggregate data rather than the behavior of individual households. Thus they believe that their estimates for movers are applicable to a housing allowance program.

This view of the adjustment process does not seem consistent either with the actual data they report or with observed market behavior. Out-of-pocket moving costs for EHAP recipients were just $54 on average in Pittsburgh and an incredible $12 in Phoenix, where only 15 percent spent anything for a moving van or even a trailer; the rest used their own vehicle, borrowed one, or somehow moved without using any at all. Nonpecuniary search costs were surely higher; households spent 3 months looking in Pittsburgh and 1 month in Phoenix. But these data strongly suggest that moving costs were low. Confirming evidence comes from the Annual Housing Survey, which reports that in the mid-1970s between 30 and 40 percent of all low-income renters in the United States moved each year.

The assumption of low moving costs does not automatically rule out Friedman and Weinberg's conclusion. It could be argued that movers quickly achieved their new equilibrium because it was easy to do so, for example; the long run may be rather short for low-income renters (at least in Phoenix). But it could also be argued that movers do not feel compelled to reach equilibrium on the first try because they can easily move again in the near future. And some of the evidence has other implications. If moving costs are so low, it is surprising that more allowance recipients did not move. Their overall mobility rate was about the same as for all low-income renters, and it appears that the allowance by itself may have induced some households to look for better housing but few to move.

A related question is whether recipients regarded their allowance payments as permanent income. The experiment had a 3-year time limit. after that, households were automatically given priority status and help in qualifying for subsidies under other programs. But, unless their unit met the quality standards of the Section 8 Existing Housing Program, which was not a requirement for subsidy in the experiment, they would have had to move or upgrade their current unit in order to continue receiving subsidy. It would be interesting to know what actually happened to the [illegible]
ment. If moving was inexpensive, many [illegible]
their original housing; if it was expens[illegible]
a 3-year allowance may not be indica[illegible]
subsidy. In the latter case, a gradual [illegible]
valid than Friedman and Weinberg co[illegible]
therefore represent more of a short-ru[illegible]
do not describe the postexperiment [illegible]

The empirical results reported in this [illegible]
of EHAP and have figured importantly in [illegible]
serious political objections to a national allowance [illegible]
be "too little" housing improvement unless rather high minimum quality standards are imposed. The low elasticity estimates in both the Demand and Supply Experiments have supported this view. Friedman and Weinberg's own policy recommendations, however, are not fully consistent with it. They initially argue that there is a serious problem of inadequate housing for low-income renters. Their work indicates that a housing allowance will not generate much improvement. But, like most economists, they conclude that the housing allowance is superior to public housing or other construction programs because of the inefficiencies inherent in in-kind transfers and the excessive cost of subsidized production. They do not suggest any program to

ameliorate the quality problem. This seems to imply that they really do not think it is very serious, after all. One wonders what set of elasticity estimates would have led Friedman and Weinberg to make different policy recommendations.

JOHN C. WEICHER

American Enterprise Institute

References

Bradbury, Katharine L., and Downs, Anthony, eds. *Do Housing Allowances Work?* Washington: Brookings Inst., 1981.

Friedman, Joseph, and Weinberg, Daniel L., eds. *The Great Housing Experiment.* Beverly Hills, Calif.: Sage, 1983.

Struyk, Raymond J., and Bendick, Marc, Jr., eds. *Housing Vouchers for the Poor.* Washington: Urban Inst. Press, 1981.

The Birth of the British Motor Car, 1769–1897 Vol. 1, *A New Machine, 1769–1842.* Vol. 2, *Revival and Defeat, 1842–93* Vol. 3, *The Last Battle, 1894–97.* By T. R. NICHOLSON.

London: Macmillan Co. (distributed in the U.S.A. by Humanities Press, Atlantic Highlands, N.J.), 1982. Pp. 163 (vol. 1); 165–333 (vol. 2), 335–506 (vol. 3). $40.00 per volume.

At first sight this book might appear to constitute low-payoff reading for an economist. It deals with the early history, indeed, really the prehistory, of the motorcar. The book's three volumes do not even bring the reader into the twentieth century, and surely the motorcar is a twentieth-century phenomenon. Moreover, the author is clearly a car buff, with more than a score of books to his credit on such subjects as vintage cars, sports cars, and even an entire volume devoted to *Car Badges of the World.* An economist who plows through the three volumes will find much that is intellectually frustrating, tedious, and repetitious. Nicholson pays practically no attention to elementary considerations of comparative cost and economic efficiency even in contexts where such considerations were obviously critical to his story and to its particular outcome. Nevertheless, economists who are interested in technical change and its political as well as economic determinants will find much that is of interest in these volumes.

This book chronicles, in extensive and loving detail, the numerous technical and commercial experiments in the application of steam to road transport. What rescues the book from pure and unabashed antiquarianism, however, is that it also pays a great deal of attention to the successive attempts of Parliament and local authorities to regulate the operation of this new transport form. Some form of road vehicle, powered by a steam engine, obviously became at least technically feasible by around 1830, at the same time that steam power was successfully adapted to the railroad. In fact, commercial steam carriages did make a brief appearance on British roads in the 1830s but, in addition to a number of other disabilities and restrictions, they encountered highly discriminatory tolls and the superior performance of the railroads for long-distance traffic.

The author discusses in great detail the attempt to deal with a new technology—the horseless carriage—by forcing it to the procrustean bed of inappropriate and sometimes totally irrelevant regulations, such as those devised for the traction engines that hauled agricultural machinery around the countryside. Under these circumstances the new steam carriages, which might eventually have exploited the considerable advantage of greater speed, were not even allowed to match the speed of a horse. In fact, a combination of private interests succeeded in imposing crippling disabilities on the operation of the newly emerging technology that essentially guaranteed its inability to compete. The attempts to repress the new transport form in nineteenth-century Britain largely succeeded, although legislative and regulatory repression must have been powerfully assisted by competition from railroads for long-distance transportation, horse-drawn trams in cities, and the technical limitations, special hazards, and unreliability of the steam carriage itself The book carefully chronicles the imposing gauntlet that the steam-powered vehicles were forced to run. The hostility of the rival stagecoach owners, landowners and farmers, turnpike trusts, and, apparently, the solid phalanx of the legal profession was expressed in discriminatory tolls, multiple license fees, restrictions on hours of operation and on access to certain urban areas, and speed limits of 4 and even 2 miles per hour. (The author, in his concluding chapter, asks, "Why did the motor car . . take root after 1896" after its earlier failures? After recovering from this singularly infelicitous metaphor, one is rather tempted to ask why more of them did not "take root" in earlier years when they were confined to top speeds of 2 miles per hour!)

Of course, from our present vantage point we might be inclined to dismiss the pre-1900 history, delineated in this book, as no more than an extended curiosum, consisting mainly of a number of British eccentrics tinkering with a steam technology that had serious, inherent limitations as a power source for small road vehicles. This new form of road transport, one might argue, was inevitably destined for only negligible success until the availability of a superior power source—the internal combustion engine—that finally established itself in the opening years of the twentieth century The appropriate response to this position would be the writing of
might give us some better sense of th
political environment less hostile than
fortunately, Nicholson offers no real g
such a counterfactual world

Certainly many contemporaries thou
much greater success in a less hostile
people," the French car manufacturer,
correspondent in 1893, "but your laws are
Tangye, a distinguished Cornish machinery [illegible]
"But for the action of a bovine Parliament, the manufacture of motor cars would have taken root in England 40 years ago, the foreign nations would have been customers instead of pioneers followed by ourselves" (p. 237)

These last two opinions were not exactly disinterested ones Nevertheless, they raise important questions. The author has provided much evidence of governmentally imposed obstructionism. Whether, in the absence of such policies, steam carriages would have progressed much more rapidly is far from clear. The book provides a valuable compendium of the nature and the sources of support for governmental obstructionism, but without equipping the reader with a real capacity for assessing its economic impact

NATHAN ROSENBERG

Stanford University

A CALL FOR PAPERS

Special Issue of the American Journal of Sociology on Sociology and Economics

Editors. Sherwin Rosen, Economics, University of Chicago, and Christopher Winship, Sociology, Northwestern University

The *AJS* announces a special issue to explore and expand the dialogue between sociology and economics in areas of mutual interest. We seek substantively oriented papers that use analytical methods to further understanding of the systematic determinants of individual behavior and the analysis of social institutions. Areas in which we believe there is significant potential for fruitful interaction include: labor markets, sorting and selection, family, education, mobility decisions, occupational choice, life-cycle analysis, fertility and demography, and the structure of organizations. Proposals for exploration of yet other areas of mutual concern are welcomed. Papers that are substantively oriented and provide original structural analysis will be given highest priority. However, papers that contrast approaches in a particular area or demonstrate the relevance and uses of work in one discipline from the other's viewpoint will be given serious consideration.

All papers will undergo the *Journal*'s standard reviewing process and should be prepared according to the "Information for Contributors" in the *AJS*. We encourage potential contributors to submit a prospectus for possible papers, especially before initiating work specifically intended for the issue. *Four copies* of the manuscript must be submitted. The relevant deadlines are listed below. Submissions should be sent to:

American Journal of Sociology
Special Issue
1130 East 59th Street
Chicago, Illinois 60637

Prospectuses should be submitted by *February 1, 1985*; papers should be submitted by *June 1, 1985*.

The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory

Alvin E. Roth

University of Pittsburgh

The organization of the labor market for medical interns and residents underwent a number of changes before taking its present form in 1951. The record of these changes and the problems that prompted them provides an unusual opportunity to study the forces at work in markets of this kind. The present paper begins with a brief history and then presents a game-theoretic analysis to explain the orderly operation and longevity of the current market, in contrast to the turmoil that characterized various earlier short-lived attempts to organize the market. An analysis is also given of some contemporary problems facing the market. A subsidiary theme of the paper concerns the history of ideas, the problems encountered in the organization of this market, and some of the solutions arrived at, anticipated the discussion of such issues [illegible] nomics and game theory

I. Introduction

This paper concerns the labor mar [illegible] dents and how it has evolved from the [illegible] present time. The paper will discuss briefly [illegible]

I have profited from conversations and correspondence on this topic with David Gale, Gene Gruver, and John S. Graettinger, M.D., as well as from the comments of participants at seminars where I have presented parts of this work over the last 4 years at Buffalo, Yale, Harvard, Northwestern, University of Pennsylvania, Virginia Polytechnic Institute, Stanford, Minnesota, and Chicago. This work has been supported by grants from the National Science Foundation and the Office of Naval Research and by fellowships from the John Simon Guggenheim Memorial Foundation and the Alfred P. Sloan Foundation.

[*Journal of Political Economy*, 1984, vol. 92, no. 6]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9206-0002$01.50

institutional procedures by which this market has been implemented and present a formal theoretical analysis of what these procedures accomplished and the problems they encountered.

From the turn of the century until 1945, the market suffered from a Prisoner's Dilemma problem in which competition by hospitals for interns manifested itself in a race to sign employment contracts earlier and earlier in a medical student's career. This problem was successfully resolved in 1945, but the market then suffered for several years from a "recontracting problem," in which the market's failure to implement an outcome in the core made it difficult for the market to clear. This recontracting problem was closely related to an "incentive problem" that put a premium on strategic behavior by market participants, which manifested itself most clearly when a centralized market mechanism was introduced in 1950. All of these problems were apparently resolved, at least in large part, by the adoption of a different centralized market mechanism in 1951, which remains in use to this day.

It will be shown that this mechanism implements an outcome in the core of the market, and some of its other properties will be studied. Some modern problems this market mechanism has encountered and the theoretical issues they raise will also be considered. One of the most interesting of these concerns the transformation of the market into one containing dual-career households, as increasing numbers of medical school graduates are married to one another and need to be employed in nearby locations. It will be shown that no core outcomes need exist in such markets.

A subsidiary theme explored in this paper concerns the history of ideas. It will be noted that the practical problems encountered by the medical profession in the organization of this labor market, and some of the solutions proposed, anticipate by a number of years the discussion of similar issues in the literature of economics and game theory.

Section II briefly discusses the history of the market, and Section III presents an analysis. Section IV analyzes some current problems facing the market. All proofs are presented in Sections V and VI, and Section VII concludes.

II. Some Institutional History

The internship was first introduced around the turn of the century as an optional form of postgraduate medical education. For students, internships offered a concentrated exposure to clinical medicine, and for hospitals they offered a supply of relatively cheap labor. The number of positions offered for interns was, from the beginning, greater than the number of graduating medical students applying for such positions,[1] and there was considerable competition among hospitals for interns.

One form in which this competition manifested itself was that hospitals attempted to set the date at which they would finalize binding agreements with interns a little earlier than their principal competitors. As a result, the date by which most internships had been finalized began to creep forward from the end of the senior year of medical school. This was regarded as costly and inefficient both by the hospitals, who had to appoint interns without knowing their final grades or class standings, and by the students and medical schools, who found that much of the senior year was disrupted by the process of seeking desirable appointments. Many resolutions were passed and much moral suasion was applied in efforts to remedy this state of affairs. A good idea of the situation in 1926 is conveyed in a letter to the Association of American Medical Colleges (Darrach 1927), the body of which is quoted in full:

> For a number of years attempts have been made to defer the appointment of hospital internes until towards the close of the fourth year. The Association of American Medical Colleges, the Council on Medical Education of the American Medical Association, and the American Hospital Association have all passed resolutions favoring this idea. The difficulty has been in persuading someone to take the lead.
>
> This is to inform you that it has been decided to defer the appointments of internes at the Presbyterian Hospital in the City of New York until some
>
> It is earnestly hoped tha
> be able to act in a similar

The advancement of the date of
less reversed, by these appeals and
an article devoted to the problem,
Darrach's letter and added that "for some
byterian Hospital soon abandoned its stand
results of a survey indicating that, in 1939, the bulk of intern appointments would be made during the Christmas holidays of the senior year, and many would be made earlier. He proposed that the problem could be solved if medical schools in the Association of American Medical Colleges (AAMC) would undertake not to give out before

---

[1] The number of positions offered for first-year medical graduates exceeded the number of eligible applicants until the mid-1970s, at which time the total of U.S. medical school graduates plus foreign graduates exceeded the number of first-year positions However, the number of positions offered continues to exceed the number of U.S. medical school graduates (see Graettinger 1976)

some agreed date any information to hospitals regarding students' abilities. This suggestion was not adopted, however.

The advancement of the date of appointment continued at an accelerating pace, and in an address to the AAMC on the problem in 1944, Turner (1945, p. 27) reported that the standard appointment date "has now been advanced on the school calendar to the *beginning of the junior year* and, indeed, inquiries now come to me even from sophomores" (emphasis added). Thus in 1944 the date of appointment had advanced to 2 full years before the internship was actually to begin.

Turner proposed a number of steps to remedy this increasingly intolerable state of affairs. Most important, he again proposed that no information about medical students should be released by medical schools before some fixed date to be agreed upon. This proposal was adopted by the AAMC in consultation with the various hospital associations, and it was decided that neither scholastic transcripts nor letters of reference would be released prior to the end of the junior year for students seeking internships commencing in 1946.

This proved to be an effective remedy for the problem it was intended to solve. Appointments for 1946 internships were largely made in the summer of 1945, and in subsequent years the dates at which information was released by medical schools was moved later into the senior year, and the date at which appointments were made followed in step. However, a new problem appeared and manifested itself in the waiting period between the time offers of internships were first made and the time students were required to accept them.

The problem was that a student who was offered an internship at, say, his third-choice hospital, and who was informed he was an alternate (i.e., on a waiting list) at his second-choice hospital, would be inclined to wait as long as possible before accepting the position he had been offered, in the hope of eventually being offered a preferable position. Students who were pressured into accepting offers before their alternate status was resolved were unhappy if they were ultimately offered a preferable position, and hospitals whose candidates waited until the last minute to reject them were unhappy if their preferred alternate candidates had in the meantime already accepted positions. Hospitals were unhappier still when a candidate who had indicated acceptance subsequently failed to fulfil his commitment after receiving a preferable offer. In response to pressure originating chiefly from the hospitals, a series of small procedural adjustments were made in the years 1945–51. The nature of these adjustments, described next, makes clear how these problems were perceived by the parties involved.

For 1945, it was resolved that hospitals should allow students 10 *days after an offer had been made to consider whether to accept or reject it.*[2] For 1946,[3] it was resolved that there should be a uniform appointment date (July 1) on which offers should be tendered (transcripts having been released on June 1) and that acceptance or rejection should not be required before July 8.[4] By 1949, the AAMC proposed ("Uniform Intern Placement Plan" 1949) that appointments should be made by telegram at 12:01 A.M., November 15, with applicants *not required to accept or reject them until 12:00 noon the same day.* Even this 12-hour waiting period was rejected by the American Hospital Association (AHA) as too long: the joint resolution ("Appointment of Interns" 1949) finally agreed on contained the phrase "no specified waiting period after 12.01 A.M. is obligatory" and specifically noted that telegrams could be filed in advance for delivery precisely at 12:01 A.M. In 1950, the resolution again included a 12-hour period for consideration, with the specific injunction that "hospitals and/or students shall not follow telegrams of offers of appointment with telephone calls" until the end of this period.[5]

By this time it was widely recognized both that there were serious problems in the last stage of the matching process and that these problems could not adequately be resolved by compressing this last stage into a shorter and shorter time period. In order to avoid these problems and the costs they imposed, it was proposed, and ultimately agreed, that a more centralized matching procedure should be tried (Mullin 1950; Mullin and Stalnaker 1951). Under this procedure, students and hospitals would continue to make contact and exchange information as before.[6] (It is worth noting in this regard that the complete job description offered by [illegible]
year was customarily specified in ad[illegible]
Thus the responsibilities, salary, etc

---

[2] Unless otherwise stated, the resolutions [illegible]
AAMC and the American Hospital Associa[illegible]
legally binding on the member hospitals [illegible]
see "Agreement on Internship Placement" (19[illegible]

[3] See "Report of the Committee on Internships" [illegible] lution it is noted that "while not observed completely by a few medical schools and hospitals, [it] had been followed by most of the institutions in the country and had resulted in a very great advance in bringing order out of a chaotic situation."

[4] "It was reported by a representative of the American Hospital Association that violations by hospitals was as high as 25 percent, nevertheless the American Hospital Association wants the system . . . to be continued" ("Internships" 1947, p. 46).

[5] "Modifications of Cooperative Plan for Appointment of Interns" (1950). Note that the injunction against telephone calls was two way, in order to stem a flood of calls both from hospitals seeking to pressure students into an immediate decision and from students seeking to convert their alternate status into a firm offer.

[6] That is, the application and interviewing process would remain unchanged. (At this time Mullin [1950] states there were about 9,300 internships being offered, to about 6,000 medical school graduates.)

ship, while they might be adjusted from year to year in response to a hospital's experience in the previous year's market, were not a subject of negotiation with individual candidates.) Students would then rank in order of preference the hospital programs to which they had applied, hospitals would similarly rank their applicants, and all parties would submit these rankings to a central bureau, which would use this information to arrange a matching of students to hospitals and inform the parties of the result.[7] A specific algorithm was proposed to produce a matching from the submitted rankings, and this will be discussed in the next section.

It was agreed to try the proposed procedure in a trial run for the 1950–51 market, which would *not* be used actually to match students and hospitals in that year. Instead, participants were asked to submit rankings as if they would be used for determining the final matching, and the plan would be evaluated for actual use after the trial run had occurred. On the basis of the trial run, the relevant medical associations agreed to adopt the procedure for the 1951–52 market. The procedure was to be voluntary: students and hospitals were both free either to participate in the process or to seek internship appointments on their own.

After the procedure was announced but before it was implemented, objections by student representatives were raised to the algorithm used to produce a matching from the rankings submitted. Specifically, they observed that this algorithm made it a matter of great importance how adroitly a student composed his rank-order list. A student might suffer if he took a "flyer" and gave high rank to hospitals he preferred but had little chance of being matched with ("The Internship Matching Plan" 1952). That is, it was noted that a student who submitted a rank order of hospitals corresponding to his true preferences might receive a less preferable match than if he had submitted a different rank order. In response to these objections, a new algorithm was substituted for the old in the 1951–52 matching plan (Mullin and Stalnaker 1952). This substitution was judged to be of sufficiently small import that its details were not widely disseminated, the announcements concerning the plan having already been distributed before the substitution was made.

In what follows, the initial algorithm will be called the "trial-run algorithm," and the algorithm that replaced it will be called the "NIMP algorithm," for National Intern Matching Program, the name adopted for this plan in 1953. The matching plan using the NIMP algorithm remains in use to this day, although the market that it serves has undergone considerable change in the intervening years.[8] This longevity and the very high levels of (voluntary) participation the system has attracted are particularly surprising in view of the turmoil in this market in the years immediately prior to the establishment of the NIMP.[9] Much of the theoretical analysis in the following section is intended to shed light on the causes of this success.

---

[7] Under the procedure being discussed here, hospitals actually ranked students in groups, rather than in a strict rank order, although in the procedure discussed next, hospitals as well as students submitted strict rank orders. Surveys of medical students (see, e.g., Rutkow and Glasgow 1978) suggest that the most important factors determining student preferences over hospital programs have to do with their reputation as educational programs, less importance being assigned to factors such as on-call schedule, patient population, and salary.

III. Theoretical Analysis

The reasons for continually advancing the date at which interns were appointed before 1945 are well understood, and little need be said about them here. Although hospitals all preferred as late an appointment date as possible, each preferred to appoint its interns earlier than its competitors: the situation is well modeled as a multiple-agent Prisoner's Dilemma. However, the negative externality produced by this competitive behavior on the part of hospitals was also experienced by the medical schools, which had no corresponding incentive to advance the appointment date. Consequently, once medical schools collectively imposed a date before which they would not release the information about students needed to make appointments, a new equilibrium was created at which there was no tendency for the date to advance.[10]

---

[8] In 1968 the National Intern Matching
Intern and Resident Matching Program
Resident Matching Program (NRMP), to
graduate medical training.

[9] In the first years, over 95 percent of
the system (cf. Mullin and Stalnaker 1952
until the early 1970s (see Checker 1973), after
reasons that will be discussed later

[10] The importance of the medical schools' acting collectively in this regard was emphasized in Curran (1945, p. 37) as follows: "While the Executive Council has already recommended that no letters of recommendation supporting internship applications should go out until after the end of the junior year, Dr. Turner's insistence that this be agreed on by all the medical schools is worthy of emphasis. Its importance may be illustrated by a situation created one year ago among the schools in New York City by such an agreement on a local basis. It was then discovered that in two neighboring cities all internship vacancies were being filled by students before the end of the third year. As a result, the New York schools were forced to abrogate hastily their rule against sending out earlier information."

A. *Stability*

To analyze the market after 1945, it will be helpful to consider it as a game whose players are a set $H = \{h_1, \ldots, h_m\}$ of hospital programs offering positions to first-year graduates, and a set $S = \{s_1, \ldots, s_n\}$ of graduating students. Each hospital program h_i has some number q_i of identical first-year positions it is seeking to fill, and each student is seeking one position.[11] The rules of the game are that any student may fill any position if and only if both the student and the hospital agree, that is, if and only if the position is proferred by the hospital and accepted by the student. Each student has preferences over all the hospital programs he is willing to accept, and each hospital program similarly has preferences over all the students it is willing to accept.[12] An agent who is not indifferent between any two acceptable alternatives (so that his preferences constitute a rank ordering) will be said to have *strict* preferences. An outcome of the game is an assignment of students to hospital programs, with no more than q_i students being assigned to any program h_i. A given outcome x is called *unstable* if some student or hospital receives an unacceptable assignment or if there exists a hospital program h_i and student s_j who each prefer the other to their assignment at x.

Formally, for each hospital program h_i, let $s_jP(h_i)s_k$ denote that h_i prefers student s_j to student s_k, let $s_jI(h_i)s_k$ denote that h_i is indifferent between the two, let $s_jP(h_i)u$ denote that h_i prefers filling a position with student s_j to leaving it unfilled, and let $uP(h_i)s_j$ denote the reverse. Then each h_i has preferences defined over its set $\{s_j \in S|s_jP(h_i)u\}$ of acceptable students. Similarly, each student s_j has preferences defined over his set $\{h_i \in H|h_iP(s_j)u\}$ of acceptable hospital programs.[13] An outcome is represented by a function $x: S \rightarrow H \cup \{u\}$, with $|\{s_j|x_j(s_j) = h_i\}| \leq q_i$.[14] For any s_j in S, $x(s_j) = h_i$ denotes that s_j is assigned a position in hospital program h_i, and $x(s_j) = u$ denotes that s_j is unassigned. For any h_i in H let $x(h_i) = \{s_j \text{ in } S | x(s_j) = h_i\}$ be the set of students assigned to h_i, it being understood that if $|x(h_i)| < q_i$ then $q_i - |x(h_i)|$ positions are unfilled in program h_i. An outcome x is unstable if for some student s_j, $uP(s_j)x(s_j)$; or if for some hospital program h_i, $uP(h_i)s_j$ for some s_j in $x(h_i)$; or if there exists a hospital program h_i and student s_j such that both (i) $h_iP(s_j)x(s_j)$ and (ii) $s_jP(h_i)\sigma$ for σ in $x(h_i)$ if $|x(h_i)| = q_i$ or σ in $x(h_i) \cup \{u\}$ if $|x(h_i)| < q_i$.

---

[11] A hospital typically offers a number of different kinds of positions, but these are separated into different programs, each administered separately.

[12] Students may regard some hospital programs as unacceptable and hospitals may so regard some students, in the sense that it would be preferable to defer taking or offering *any* appointment rather than one that is unacceptable. For example, students and hospitals who have not exchanged the necessary preliminary information (e.g., transcripts and interviews) may be expected to regard one another as unacceptable (at least for the time being). Since interviews are time consuming, students face a problem in deciding what hospitals to apply to, and hospitals face a similar problem in deciding whom to invite for interviews, but these problems will not be addressed here.

[13] We will not need to be concerned over what an agent's preferences may be over those alternatives that are unacceptable. However, the formal description of the game given here is incomplete in another respect as well, since the preferences of hospitals *for groups of students have not been specified* (see Roth 1984*b*). The description given here will nevertheless be complete enough for the purposes of this paper.

[14] For any set T, $|T|$ denotes the number of elements in T, so, e.g., $|\{s_j|x(s_j) = h_i\}|$ is the *number of students employed by* h_i *at the outcome* x

Thus an outcome is unstable if it gives some student or hospital program an unacceptable (not individually rational) assignment or if there exists a student s_j and hospital program h_i such that (i) s_j prefers h_i to the hospital $x(s_j)$ and (ii) h_i prefers s_j to the incumbent in one of its q_i available positions. When no confusion will result, a hospital program will henceforth sometimes be referred to simply as a hospital.

An outcome that is not unstable in this sense is called *stable*, and it is straightforward to verify that the set of stable outcomes is the core defined by weak domination of this game.[15] Denote the set of stable outcomes by $C(\mathbf{P})$, where $\mathbf{P}$ is the $(m + n)$-tuple of agent preferences: that is, $\mathbf{P} = [\mathbf{P}(h_1), \ldots, \mathbf{P}(h_m), \mathbf{P}(s_1), \ldots, \mathbf{P}(s_n)]$. The vector $\mathbf{P}$ will sometimes be called the *preference profile*.

Consider now a set of job offers from hospitals to (acceptable) students, which, if the students each accept the best of the offers they have received (including the possibility of remaining unmatched), would result in an unstable outcome x. The fact that x is unstable means that there is a hospital h_i and a student s_j who would both prefer to x another outcome at which s_j was employed by h_i. So these two agents have an incentive to try to locate [illegible] and we might expect to witness the kind of last-[illegible] intern market prior to 1951. Look[illegible] might expect that any voluntary [illegible] would experience similar turmoil [illegible] Since the NIMP is a voluntary s[illegible] degree of orderly participation [illegible] conjecture that it produces stable outcomes, that is, outcomes in the core of the market. The following theorem, which will be proved in Section V, supports this conjecture.

---

[15] For markets with side payments, the core de[illegible] with the core defined by strong domination, and in the market studied here when [illegible] 1 for each hospital h_i, the two cores also coincide. However, in the more general case considered here, the core defined by weak domination is contained in the core defined by strong domination. See Roth and Postlewaite (1977) for a related discussion. For our purposes, the distinction will not be consequential, and the set of stable outcomes will henceforth be referred to simply as the core, and stable outcomes will also be called core outcomes. However, there are differences between the special case when all $q_i = 1$ and the general case considered here that bear on this distinction (see Roth 1984*b*).

[16] A system that produces unstable outcomes gives at least some agents the incentive to go outside the system to find one another and arrange their employment. Only in a system where participation is voluntary, however, can we expect to observe agents acting on these incentives.

THEOREM 1: Let **P** be the vector of preferences (the rank-order lists) submitted by the students and hospitals. Then the NIMP algorithm produces an outcome in $C(\mathbf{P})$.

When we consider in Section III*B* the incentives agents have to state their true preferences, we will examine from another viewpoint the question whether a stable outcome is achieved under the current organization of the market. First, however, we consider some results that concern the fact that different outcomes in the core differentially favor the two sides of the market. Specifically, theorem 2 states that, in a certain strong sense, the outcome produced by the NIMP algorithm is the best core outcome for all the hospitals and the worst for all the students. The following definitions will be used in stating the theorem.

For a given preference profile **P**, a student s_j and a hospital h_i will be called *achievable* for one another if there is some stable outcome at which s_j is employed by h_i, that is, if there is an outcome x in $C(\mathbf{P})$ such that $x(s_j) = h_i$. For each h_i, let p_i be the number of achievable students; that is, $p_i = |\{s_j | s_j \text{ is possible for } h_i\}|$, and define k_i by $k_i = \min\{q_i, p_i\}$. Thus k_i equals the minimum of the number of positions offered by h_i or the number of different students who could fill those positions at some stable outcome.

THEOREM 2: Let **P** be the preference profile (i.e., the rank-order lists) submitted by the students and hospitals. Then the outcome produced by the NIMP algorithm has the property that each hospital h_i is assigned its k_i highest ranked achievable students, and each student s_j is assigned his lowest ranked achievable alternative.[17]

Note that not only is there no trade-off between hospitals—they all simultaneously receive their best achievable students—but within a hospital there need be no trade-off between achievable students. The theorem implies, for example, that if a hospital h_i with two positions ($q_i = 2$) employs its first- and third-ranked student at one stable outcome and its second- and fourth-ranked student at another stable outcome, then at the outcome produced by the NIMP algorithm it is assigned its first- and second-ranked student. The theorem is proved in Section V.

Note that theorems 1 and 2 can both be stated as existence results. Theorem 1 states that the core of the market is nonempty, and theorem 2 states that (when the preference profile consists of rank orderings—i.e., strict preferences) there is a core outcome that is best for all the hospitals and worst for all the students. These existence results were first stated by Gale and Shapley (1962), who were unaware of the NIMP procedure, which had then been in operation for 10 years.[18] In order to prove these existence results, they developed an algorithm that is considerably simpler than the NIMP algorithm but turns out to be equivalent; that is, for a given preference profile of strict preferences, the two algorithms produce the same stable outcome.[19] These existence results have recently been shown to hold in considerably more general kinds of labor markets (see Kelso and Crawford 1982; Roth 1984*a*, 1984*d*).

---

[17] For a student (or hospital) who is unmatched at some stable outcome, being unmatched will be the lowest-ranked achievable alternative.

A corollary of the proof of theorem 2 is that there exists a stable outcome that is optimal for students in the same sense that the outcome described in theorem 2 is optimal for hospitals. That is, we can state the following:

COROLLARY 2.1: For any profile **P** of strict preferences, there exists a stable outcome with the property that each student s_j is assigned his highest-ranked achievable alternative.

B. Incentives

Let us now consider the incentives faced by agents when a centralized market mechanism is employed that requires as input agents' stated preferences over potential assignments. The adoption of any such mechanism creates a new game among the agents, who must decide what preferences to state. This point can perhaps be made most clearly by briefly considering the trial-run algorithm that was discarded in favor of the NIMP algorithm
lems it was observed to cause.

Under the trial-run procedure
students submitted a rank orderi
applied, and hospitals submitted a
five groups. The hospital gave a
students, equal in size to the numb

---

[18] As a historical note, I am indebted to David Gale
aware of the labor market for interns in 1976, when he was told of it by a physician who heard him speak on the "marriage problem." At that time, he sent a copy of Gale and Shapley (1962) to an administrator of the NIMP, which seems to have been the first time that anyone associated with the program became aware of the game-theoretic formulation of the problem and the results concerning optimal outcomes in the core

[19] Gale and Shapley (1962) assumed that all agents had strict preferences. The NIMP algorithm, since it requires its input to be rank orderings, requires all agents to state their preferences as if they were strict—i.e., no indifferences can be expressed. Theorem 1 and the existence of stable outcomes are unchanged when agents are indifferent between some alternatives, but we will see in Sec. IV that theorem 2 becomes an artifact.

TABLE 1

	Agent					
Rank	s_1	s_2	s_3	h_1	h_2	h_3
1	h_1	h_2	h_1	s_2	s_1	s_3
2	h_2	h_3	h_3	s_3	s_2	s_2
3	h_3	h_1	h_2	s_1	s_3	s_1

larly it gave a rank of 2 to an equal-size group of second-choice students, ranked less desirable students 3 and 4, and ranked unacceptable students 5. The matching algorithm proceeded in stages, or runs. At the first of these, called the 1:1 stage, students and hospitals were matched if they had each given the other a rank of 1. The next stage was a 1:2 stage, where remaining students and hospitals were matched if the student had ranked the hospital 1 and the hospital ranked the student 2. Then followed a 2:1 stage, a 2:2 stage, a 1:3 stage, and so forth.

The difficulty observed by the student representatives after the trial run was that a student who did not match with his first-choice hospital might find that, say, his second-choice hospital had no places left by the time the 2:1 stage rolled around, so that even if his second choice had ranked him a 1, he might end up assigned to a lower-choice hospital. In this case, the student would have done better to have submitted a different rank ordering. The following example, in which there are three students and three hospitals, each offering one position, will make this completely clear. The rank orderings given in table 1 correspond to the true preferences of the agents.

If the agents all submit their true rank orderings, the trial-run algorithm finds no matches at the 1:1 stage, and at the 1:2 stage it matches s_2 with h_2 and s_3 with h_1, so that s_1 is ultimately matched with h_3, his third choice.[20] However, if s_1 had ranked his second choice h_2 first, he would have been matched with it (at the 1:1 stage). So s_1 has an incentive to misrepresent his preferences strategically when submitting his rank order.[21]

---

[20] Note that this outcome is not stable. With respect to the preferences of the players as given in table 1, there are exactly two stable outcomes, at both of which s_1 is matched with h_2. So the hospital-optimal stable outcome x is given by $x(s_1) = h_2$, $x(s_2) = h_1$, $x(s_3) = h_3$, which gives every hospital its first-choice student. The student-optimal stable outcome y is given by $y(s_1) = h_2$, $y(s_2) = h_3$, $y(s_3) = h_1$, which gives hospitals h_1 and h_3 their second-choice student.

[21] A very similar matching procedure, which exhibits precisely the same incentive problem, is currently employed by the Fulbright Senior Scholars Awards Program, which matches American scholars with foreign universities (see Council for International Exchange of Scholars 1984). I am indebted to Gene Gruver for pointing this out to me.

The NIMP algorithm was introduced with the claim that it solved this problem (see Mullin and Stalnaker 1952; Graettinger and Peranson 1981). Under the NIMP procedure, both students and hospitals submitted rank orderings. It was essentially claimed that the NIMP algorithm made it a dominant strategy for every agent to state his true preferences; that is, each agent could do no better than to submit a rank ordering corresponding to his true preferences,[22] no matter what preferences were submitted by other agents. The following results demonstrate that this claim is not quite correct, but that the NIMP algorithm comes as close in a certain sense to fulfilling this claim as is possible for a procedure that produces a stable outcome.

Theorems 3–7 will be stated here without proofs, which can be found in Roth (1982*a*) (theorems 3–6) and Roth (1984*b*, 1984*c*) (theorem 7). The first of these results is an impossibility theorem for all stable matching procedures, which are defined to be functions from preference profiles **P** to stable outcomes $x(\mathbf{P})$ in $C(\mathbf{P})$.

THEOREM 3: No stable matching procedure exists for which it is a dominant strategy for all agents to state their true preferences.

Since the NIMP is a stable matching procedure, this theorem implies that there are circumstances in which it gives some agents the incentive to submit a rank ordering different from their true preferences. Specifically, the proof of this theorem in Roth (1982*a*) implies that it is not a dominant strategy for students participating in the NIMP to state their true preferences. However, the following results show that the NIMP procedure offers some relief from the incentive problems present in the trial-run algorithm that it replaced.[23]

THEOREM 4: The NIMP procedure giv [illegible] incentive
to misrepresent his true first cho

In particular, any student who
element differs from his true fi
well by submitting an otherwise
element is his true first choice
that the NIMP procedure contra
algorithm.

---

[22] Let **P*** be some agent's true preference ordering, and let **P** be the rank ordering he submits, which is necessarily a strict ordering. If the true preferences **P*** are also strict then **P** corresponds to **P*** if and only if **P** = **P***. If the true preferences **P*** are not strict then **P** corresponds to **P*** if and only if the fact that one alternative is strictly preferred to another under **P*** implies it is strictly preferred under **P**. (So if **P*** expresses indifference between two alternatives, either may be ranked higher under **P** corresponding to **P***.)

[23] In Roth (1982*a*, 1984*b*), theorems 4, 5, and 7 were stated in terms of a stable matching procedure that always yielded the optimal stable outcome for a given side of the market. From the point of view of incentives, any two such procedures are equivalent. Here, therefore, the theorems can be stated in terms of the NIMP procedure, which by theorems 1 and 2 yields the hospital-optimal stable outcome.

Since students are each assigned to only one hospital, the model presented here is adequate for discussing their incentives, since their preferences are defined over hospitals and their strategies involve rank orderings of hospitals. However, in order to discuss adequately the incentives facing hospitals that have more than one position to fill, it would be necessary to specify their preferences over groups of students. A complete description of the general case would take us well beyond the scope of this paper. But for the simple case of hospitals with only one position to fill ($q_i = 1$), the model discussed here is complete enough to allow us to state the following.

THEOREM 5: The NIMP procedure makes it a dominant strategy for all hospitals with one position to state their true preferences.

The NIMP procedure thus avoids some of the simple opportunities for misrepresentation offered by the trial-run algorithm. No student can profit from misrepresenting his first choice, and in the simple case in which hospitals have only a single position to offer, it is a dominant strategy for hospitals to state their true preferences.[24] The following result shows that, in a certain sense, the NIMP procedure does as well as possible in this regard; no stable procedure exists that does all this and that also never makes it profitable for a student to misrepresent his second choice (even in the simple case in which hospitals have only single positions).

THEOREM 6: No stable matching procedure exists for which it is a dominant strategy for all agents to state their true kth choice, for $k \neq 1$.

In view of the fact that the NIMP procedure does not make it a dominant strategy for students to state their true preferences, we must now reconsider what reason there is to believe that the NIMP procedure in practice yields a stable outcome (in terms of the true preferences of the agents). That is, theorem 1 states that, for any preference profile **P** of submitted rank-order lists, the NIMP procedure yields an outcome in $C(\mathbf{P})$, that is, a stable outcome with respect to the *stated* preferences. But since some students may have an incentive to misrepresent their preferences, the stated preferences **P** may not coincide with the true preferences **P***. What reason, then, do we have to believe that the outcome produced will be stable with respect to the true preferences? To put it another way, how can we explain the empirical fact that the NIMP procedure has produced an orderly market since 1951, in contrast to the turmoil associated with the instability characteristic of the earlier forms of market organization it replaced?

---

[24] Theorem 5 is included here to help illustrate the manner in which the NIMP procedure comes close to achieving its desired aim of giving neither students nor hospitals any incentive to misrepresent their preferences. However, it is shown in Roth (1984*b*) that the NIMP procedure does not in general make it a dominant strategy for hospital programs having more than one position to state their true preferences, and that in fact no stable matching procedure exists that has this property for such hospital programs.

Two kinds of answers can be given to this, depending on how much information the agents possess. First, if agents have little information about the preferences of other agents, they will not in general be able to determine an "optimal" misrepresentation. Together with the fact that the literature distributed to students and hospitals with the forms for their rank-order lists continues to claim that it is optimal to submit true preferences, this may persuade them to submit their true preferences. If so, the resulting outcome would be stable.

However, there is at least anecdotal evidence that many students are *not* persuaded that it is optimal to submit their true preferences, and there is substantial information available to them about the historical preferences of hospitals and students (e.g., which hospitals are prestigious, and what kind of medical-school records their interns must typically have to get an appointment). We must therefore consider whether it is consistent to expect that a stable outcome will result from the NIMP procedure when agents are rational and well informed, that is, when stated preferences can be expected to differ from true preferences. The traditional formulation in game theory of how such agents behave is that their stated preferences (i.e., strategy choices) will constitute a Nash equilibrium, so that each agent's stated preference gets him the best match obtainable given the stated preferences of the others. The following theor
tent to expect, in this case also, that th
a stable outcome.

THEOREM 7: There exist Nash equ
lists such that the outcome produced
with respect not only to **P**, but also t
the NIMP outcome is contained in

IV. Contemporary Issues

We now turn to some contemporary issues that have been raised concerning the present operation of the National Resident Matching Program (NRMP). In a certain sense, each of these issues concerns the underlying demographics of the market.

When the NIMP algorithm was first employed, there were many more internships being offered than there were students to fill them, and consequently all core outcomes were rather favorable for students. By the mid-1970s, however, the inclusion of foreign medical school graduates in the matching program had succeeded in closing the gap between the number of applicants and positions offered, and outcomes in the core were no longer so favorable to students.[25] It is thus not surprising that the question of whether the NIMP algorithm unfairly favors hospitals at the expense of students eventually became an issue.[26] However, the following result shows that the conflict of interest between students as a group and hospitals as a group evidenced in theorem 2 is at least partly an artifact of the NIMP procedure's requirement that participants submit rank orders (i.e., strict preferences) even when they may be indifferent between some potential matches.

THEOREM 8: When preferences are not all strict, there may exist no two stable outcomes such that all hospitals prefer one and all students prefer the other.

Thus when preferences are not strict, it is not in general meaningful to speak of a procedure that favors hospitals over students. In the example used in the proof of this theorem, in Section VI, any stable procedure must favor some hospitals and students at the expense of other hospitals and students.

Another issue that has been raised concerns the geographic distribution of interns and residents to hospitals. For example, Sudarshan and Zisook (1981) write:[27]

> Williams and his colleagues correctly point out that a great deal of subtlety and structure in the present matching system are probably not appreciated by the casual observer. Nevertheless, an obvious answer to their question of why the present system and not its mirror image holds sway is apparent. The United States suffers a terrible problem of maldistribution of physicians, with urban areas being relatively over-

---

[25] Mullin and Stalnaker (1952) report that in the first year of operation of the NIMP procedure, 10,414 positions were offered through the program to 5,681 student participants, of whom 98 percent were assigned internships by the algorithm. (They further report that of these students, 84 percent received their highest-ranked choice and another 10 percent their second choice.) By 1976, Graettinger (1976) reports that 16,112 positions were offered through the program, compared to 16,728 student participants (of whom 11,735 were U.S. medical school graduates). In that year only 73 percent of the students (including 92 percent of the U.S. students) were matched by the algorithm. (Each year students and hospitals who are unmatched are able to seek one another out after the matching procedure has taken place.)

[26] See Williams, Werth, and Wolff (1981), who reach a conclusion resembling theorem 2 on the basis of some examples. By this time, Gale and Shapley's 1962 paper is sometimes referenced in the medical literature (see n. 18), not always correctly. However, a number of authors appeared to be aware of the equivalence of the NIMP algorithm and the Gale-Shapley algorithm (in at least some cases reaching this conclusion because of a belief that the two algorithms are the same). In Sec. VII, the difficulty of precisely tracing the history of ideas in the literature will be briefly discussed.

[27] Sudarshan and Zisook (1981) wrote in response to Williams et al. (1981). They were aware of Gale and Shapley (1962), but believed it implied that only two stable outcomes exist, namely a hospital-optimal stable outcome and its "mirror-image" student-optimal outcome.

> served and inner-city and rural areas being relatively underserved. At present, although approximately 100 hospitals fill every residency position, there are over 100 hospitals that do not receive a single application. This maldistribution would only be worsened by the only rational alternative to the present matching program, the mirror-image program, which favors students. [P. 526]

The following theorem shows that, in fact, when all preferences are strict (so that hospital and student optimal stable outcomes exist), the selection of a particular stable outcome will affect only which students are assigned to which hospitals and will not have any effect on which hospital positions are filled and which students are matched. That is, the particular stable outcome selected will not have any influence on the geographic distribution of matched students.

THEOREM 9: When all preferences are strict, the set of hospital positions filled is the same at every stable outcome, as is the set of students who are assigned positions

This theorem is proved in Section VI.

The final issue to be considered here concerns the fact that increasing numbers of medical students marry other medical students and seek to be assigned positions in the same community. At present these couples have two options: they may enter the centralized matching process together and be matched by means of a special "couples algorithm," or they may elect to remain outside of the centralized process and negotiate directly with hospital pr [illegible] increasing numbers of couples hav [illegible] clined to participate in the match, a [illegible] positions with hospitals.[28]

In the light of the preceding an [illegible] couples and hospitals both find it p [illegible] ments outside of the matching progr [illegible] algorithm produces outcomes that are uns [illegible] ples. This turns out to be the case The couples alg [illegible] until 1983 did not permit a couple to specify its preferences over pairs of positions. Instead, each couple submitted a rank-order list for each spouse and assigned one spouse priority in the match. Elias and Elias (1980) discuss some of the drawbacks of this couples algorithm and propose that it be replaced by one allowing each couple to submit a rank-order list expressing its preferences over pairs of positions. (In the 1983 match, couples were for the first time allowed to express such preferences; see the *NRMP Directory* [1983].) The following result indicates, however, that even when couples are allowed to state their preferences more accurately, the problem of finding a stable outcome may still be intractable, since such an outcome need not exist.

---

[28] Quite early, Checker (1973) attributed the declining rate of participation in the matching program partly to married couples (Hospitals that negotiate directly with couples may still seek to fill their remaining positions through the matching program)

THEOREM 10: In a market in which some agents are couples, the set of stable outcomes may be empty.

The proof is contained in Section VI.

V. The NIMP Algorithm and Proofs of Theorems 1 and 2

The NIMP algorithm (see Stalnaker 1953; Darley 1959; *NIRMP Directory* 1979) works as follows. Each hospital program rank orders the students who have applied to it (marking "X" any students who are unacceptable) and each student rank orders the hospital programs to which he has applied (similarly indicating any that are unacceptable). These lists are mailed to the central clearinghouse, where they are edited by removing from each hospital program's rank-order list any student who has marked that program as unacceptable and by removing from each student's list any hospital that has indicated he is unacceptable. (I am indebted to John S. Graettinger for explaining this initial editing.) The edited lists are thus rank orderings of acceptable alternatives.

These lists are entered into what may be thought of as a list-processing algorithm consisting of a matching phase and a tentative-assignment-and-update phase. The first step of the matching phase (the 1 : 1 step) checks to see if there are any students and hospital programs that are top ranked in one another's ranking. (If a hospital h_i has a quota of q_i, then the q_i highest students in its ranking are top ranked.) If no such matches are found, the matching phase proceeds to the 2 : 1 step, at which the second-ranked hospital program on each student's ranking is compared with the top-ranked students on that hospital's ranking. At any step, when no matches are found the algorithm proceeds to the next step, so the generic k: 1 step of the matching phase seeks to find student-hospital pairs such that the student is top ranked on the hospital's ranking and the hospital is kth ranked by the student. At any step where such matches are found, the algorithm proceeds to the tentative-assignment-and-update phase.

When the algorithm enters the tentative-assignment-and-update phase from the k: 1 step of the matching phase, the k: 1 matches are tentatively made; that is, each student who is a top-ranked choice of his kth-choice hospital is tentatively assigned to that hospital. The rankings of the students and hospitals are then updated in the following way. Any hospital that a student s_j ranks lower than his tentative assignment is deleted from his ranking (so the updated ranking of a student s_j tentatively assigned to his *k*th choice now lists only his first *k* choices), and student s_j is deleted from the ranking of any hospital that was deleted from s_j's ranking (so the updated rankings of each hospital now include only those applicants who have not yet been tentatively assigned to a hospital they prefer). Note that if one of a hospital's top-ranked candidates is deleted from its ranking, then a lower-ranked choice moves into the top-ranked category, since the hospital's updated ranking has fewer students, but the same quota, as its original ranking. When the rankings have been updated in this way, the algorithm returns to the start of the matching phase, which examines the updated rankings for new matches. Any new tentative matches found in the matching phase replace prior tentative matches involving the same student. (Note that new tentative matches can only improve a student's tentative assignment, since all lower-ranked hospitals have been deleted from his ranking.) The algorithm terminates when no new tentative matches are found, at which point tentative matches become final. That is, the algorithm matches students with the hospitals to which they are tentatively matched when the algorithm terminates. Any student or hospital position that was not tentatively matched during the algorithm is left unassigned and must make subsequent arrangements by negotiating directly with other unmatched students or hospitals. (See fig. 1 for a schematic of the algorithm.)

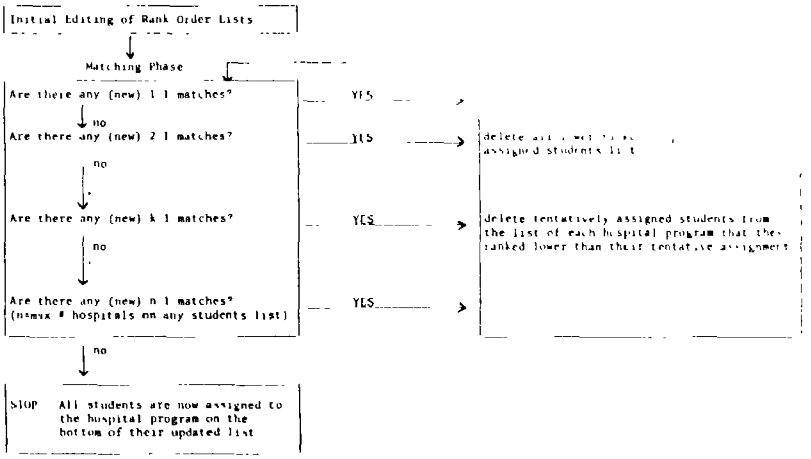

Initial Editing of Rank Order Lists
Matching Phase
Are there any (new) 1 1 matches?
no
Are there any (new) 2 1 matches?
no
Are there any (new) k 1 matches?
no
Are there any (new) n 1 matches?
(n=max # hospitals on any students list)
no
YES
YES
YES
YES
delete all [illegible]
assigned students list
delete tentatively assigned students from the list of each hospital program that they ranked lower than their tentative assignment
STOP All students are now assigned to the hospital program on the bottom of their updated list

FIG. 1.—The NIMP algorithm

The proofs of theorems 1 and 2 given below closely follow the treatment of Gale and Shapley (1962) and together amount to a formal proof that the NIMP algorithm and the algorithm considered by Gale and Shapley are equivalent.

PROOF OF THEOREM 1: When the algorithm terminates, each hospital h_i is matched with the top q_i choices (i.e., the top-ranked choices) on its final updated rank-order list. (This follows, since the algorithm does not terminate while tentative k:1 matches can still be found.) This assignment is stable, since any student s_j whom some hospital h_i originally ranked higher than one of its final assignees was deleted from h_i's ranking when s_j was given a tentative assignment higher in his ranking than h_i. Hence the final assignment gives s_j a position he ranked higher than he ranked h_i. So the final outcome is not unstable with respect to any such h_i and s_j.

PROOF OF THEOREM 2: To prove the first part of the theorem, it will be sufficient to show that no achievable student is ever deleted from a hospital program's rank-order list, so that the final assignment gives each hospital program its best achievable assignments. This can be seen by induction. Suppose that, up to the rth iteration of the algorithm, no student has been deleted from the ranking of a hospital for which he is achievable, and that on the $(r + 1)$st iteration student s_j is tentatively matched with hospital h_i and deleted from the ranking of hospital h_k. Then any assignment that matches s_j with h_k and assigns achievable matches to h_i is unstable, since s_j ranked h_i higher than h_k and h_i ranked s_j higher than at least one of its assignees. (This follows since s_j was top ranked by h_i at the end of the rth iteration, when no achievable students had yet been deleted from h_i's rank-order list.) So s_j is not achievable for h_k.

The second part of the theorem is now straightforward, since any outcome y that some s_j likes less than the hospital-optimal outcome x is unstable if the hospital $x(s_j)$ is assigned some student at y whom it likes less than s_j.

Corollary 2.1 follows in a straightforward way by considering the NIMP algorithm with the roles of students and hospitals reversed, so that each student is treated as a hospital with one position to offer and each hospital is treated as a student with q_i top-ranked choices.

VI. Proofs of Theorems 8, 9, and 10

PROOF OF THEOREM 8: Consider the market in which there are three hospitals, $H = \{h_1, h_2, h_3\}$, each offering one position, and three students, $S = \{s_1, s_2, s_3\}$. Hospital h_1 is indifferent between s_1 and s_2, whom it prefers to s_3, that is, $s_1 I(h_1) s_2 \mathbf{P}(h_1) s_3$. Hospital h_2 strictly prefers s_1 to s_3 to s_2, while h_3 prefers s_2 to s_3 to s_1. The preferences of the students are given by $h_1\mathbf{P}(s_1)h_2\mathbf{P}(s_1)h_3$, $h_1\mathbf{P}(s_2)h_3\mathbf{P}(s_2)h_2$, and $h_1\mathbf{P}(s_3)h_2\mathbf{P}(s_3)h_3$. There are exactly two stable outcomes in this market, x and y, given by $x(s_1) = h_1, x(s_2) = h_3, x(s_3) = h_2, x(h_1) = s_1, x(h_2) = s_3, x(h_3) = s_2$, and $y(s_1) = h_2, y(s_2) = h_1, y(s_3) = h_3, y(h_1) = s_2, y(h_2) = s_1, y(h_3) = s_3$. Hospital h_1 is indifferent between its assignment at x and at y, while hospital h_2 and student s_2 prefer y to x, and hospital h_3 and students s_1 and s_3 prefer x to y. So each stable outcome is preferred to the other by some students and some hospitals.

Proof of Theorem 9: For any outcome z, let $\mathbf{P}(z)$ denote the set of hospital positions filled at z. (To make the definition precise, let the q_i identical positions offered by hospital program h_i be numbered, so that if k of these positions are filled at z they are taken to be the *first* k positions.) Let $p(z) = |\mathbf{P}(z)|$ be the number of positions filled. Similarly let $Q(z)$ denote the set of students assigned to positions at outcome z, and let $q(z)$ denote their number (so $n - q[z]$ students are unmatched at z). Let x be the hospital-optimal stable outcome, and let y be the student-optimal stable outcome; both exist since all preferences are strict. Then theorem 2 implies that $\mathbf{P}(x)$ contains $\mathbf{P}(z)$ for all stable outcomes z, and consequently $p(x) \geqslant p(z)$. This latter inequality implies $q(x) \geqslant q(z)$ for all stable outcomes z, since every position is filled by exactly one student, that is, since $p(z) = q(z)$ for every outcome z. Theorem 2 also implies that $Q(x)$ is contained in $Q(z)$ for all stable outcomes z, so $q(x) \leqslant q(z)$. Hence $Q(x) = Q(z)$ for all stable z: that is, the set of students who are assigned positions is equal to $Q(x)$ at every stable outcome z. But $q(x) = q(z)$, and hence $p(x) = p(z)$ for every stable outcome z. Since $\mathbf{P}(x)$ contains $\mathbf{P}$

the set of hospital positions filled at

$\mathbf{P}(x)$.

Proof of Theorem 10: Consider th

$h_2, h_3, h_4\}$, each of which offers exact

has strict preferences over students

The students consist of two married cou

couple has strict preferences over ordered pairs of hospitals, as given in table 2. Thus couple $\{1, 2\}$ have as their first choice that student 1 be matched with h_1 and student 2 with h_2, and have as their last choice that student 1 be matched with h_2 and student 2 with h_1. The 24 possible assignments of students to hospitals are listed in table 3, along with the reason each such outcome is unstable. Thus outcome 1, which assigns student i to hospital h_i, $i = 1, \ldots, 4$, is unstable because both hospital h_2 and couple $\{3, 4\}$ would prefer that student 4 be matched with h_2. (This follows since h_2 prefers student 4 to student 2, and $\{3, 4\}$ prefers h_3h_2 to h_3h_1.) Note that the emptiness of the set of stable outcomes here is not a "knife-edge" phenomenon. The example would be completely unchanged, for instance, if any preferences for h_2 that kept student 4 as the first choice were substituted

TABLE 2

PREFERENCES OF HOSPITALS AND COUPLES

HOSPITALS' RANK ORDERS				COUPLES' RANK ORDERS	
h_1	h_2	h_3	h_4	{1, 2}	{3, 4}
4	4	2	2	h_1h_2	h_4h_2
2	3	3	4	h_4h_1	h_4h_3
1	2	1	1	h_4h_3	h_4h_1
3	1	4	3	h_4h_2	h_3h_1
				h_1h_4	h_3h_2
				h_1h_3	h_3h_4
				h_3h_4	h_2h_4
				h_3h_1	h_2h_1
				h_3h_2	h_2h_3
				h_2h_3	h_1h_2
				h_2h_4	h_1h_4
				h_2h_1	h_1h_3

TABLE 3

EVERY OUTCOME IS UNSTABLE

Outcome	h_1	h_2	h_3	h_4	Unstable with respect to.
1	1	2	3	4	4, h_2
2	1	2	4	3	4, h_2
3	1	3	2	4	2, h_1
4	1	3	4	2	4, h_1
5	1	4	2	3	2, h_4
6	1	4	3	2	4, h_1
7	2	1	3	4	4, h_1
8	2	1	4	3	4, h_2
9	2	3	1	4	2, h_3
10	2	3	4	1	4, h_1
11	2	4	1	3	2, h_1
12	2	4	3	1	4, h_1
13	3	1	2	4	4, h_2
14	3	1	4	2	2, h_3
15	3	2	1	4	2, h_4
16	3	2	4	1	2, h_3
17	3	4	1	2	1, h_1
18	3	4	2	1	2, h_1
19	4	1	2	3	4, h_2
20	4	1	3	2	2, h_3
21	4	2	1	3	2, h_4
22	4	2	3	1	2, h_3
23	4	3	1	2	3, h_3
24	4	3	2	1	4, h_4

VII. Concluding Remarks

The NIMP algorithm, which was adopted in 1951, produces the hospital-optimal outcome in the core of the market. It is therefore noteworthy that the core was not formally defined as an independent solution concept in game theory until 2 years later, in the work of Gillies (1953*a*, 1953*b*) and Shapley (1953). The fact that the core was so structured in such markets as to contain optimal outcomes of this kind was not known until the work of Gale and Shapley (1962). The kinds of incentive questions that prompted the replacement of the 1950 trial-run algorithm were not commonly discussed in economics until at least 10 years later (see Green and Laffont [1979] for a survey), and the specific incentive properties of stable procedures in markets of this kind have been well understood only in the last few years.[29] It is difficult to say precisely what was known about the NIMP algorithm when it was implemented, since the associated literature contains a number of correct assertions about its properties, alongside other assertions that are incorrect. However, it is clear that several ideas that subsequently became important in economic theory were earlier encountered in the course of organizing this medical labor market.[30]

A number of the contemporary problems arising in the medical labor market raise theoretical questions that seem likely to have more general significance. Perhaps the most topical of these is the problem of two-career couples referred to in the previous section. Since theorem 10 applies to quite general kinds of labor markets, some of the changes caused by increasing numbers of dual-career households in other labor markets may prove related [illegible]

---

[29] There has been a considerable amoun [illegible] and other questions concerning matching [illegible] existence of stable outcomes is given by Q [illegible] including both marriage markets and exch [illegible] kind studied in Shapley and Scarf (197 [illegible] (1982*b*). Theorem 3 has been independe [illegible] (1982), a strengthening of theorem 5 to includ [illegible] and Freedman (1981) and Ritz (1982), and Gale an [illegible] vation analogous to theorem 9. Similar results are explored [illegible] (1983) for a related model of matching, and I am indebted to David Gale for informing me that he and M. Sotomayor have obtained some results similar to those in Sec IV. The structure of the set of stable outcomes in general markets of this kind is explored in Roth (1984*a*).

[30] One of the goals expressed for the NIMP procedure was that it should achieve the same outcome as an orderly free market (see the brief historical review in Graettinger and Peranson 1981) Since it achieves an outcome in the core, there is a well-defined sense in which this goal is achieved. Partly because of the two-sided nature of labor markets, such a procedure is quite different from similarly motivated procedures for one-sided markets, as when students are to be assigned to dormitories (see Hylland and Zeckhauser 1979; Leonard 1983).

Another contemporary problem concerns adapting the matching procedure to the market for advanced residency positions, which are less homogeneous than positions for first-year graduates.[31] And there remain open questions about the market described here and the forces prompting the adoption of this particular form of market organization. Specifically, the precise role that salary adjustments play in this market is still unclear, as are the market features that prevented salaries alone from clearing the market in an orderly way in the late 1940s.[32]

In conclusion, it bears repeating that the history of rule changes a market has undergone provides a valuable perspective from which to view the forces at work in the market. Game theory, which is the part of economic theory most explicitly concerned with the implications of different rules of organization, provides a natural theoretical framework to study these changes.

References

"Agreement on Internship Placement." *J. Assoc. American Medical Colleges* 20 (May 1945): 192–94.

"Appointment of Interns." *J. Assoc. American Medical Colleges* 24 (May 1949): 174–75.

Bergstrom, Theodore C., and Manning, Richard. "Can Courtship Be Cheatproof?" Mimeographed. Ann Arbor: Univ. Michigan, 1982.

Checker, Armand. "The National Intern and Resident Matching Program, 1966–72." *J. Medical Educ.* 18 (January 1973): 106–9.

Council for International Exchange of Scholars. *Fulbright Senior Scholars Awards Program, 1984/85.* Washington: Council for Internat. Exchange of Scholars, 1984.

Curran, J. A. "Letter." *J. Assoc. American Medical Schools* 20 (January 1945): 37.

---

[31] The labor market for advanced residencies is experiencing difficulties at present similar in origin to those experienced by the market for interns in the period covered here. Now, as then, the medical literature discussing the problem and what to do about it reflects a division of opinion between those who attribute the problems to flaws in the organization of the market and those who attribute them to flaws in the moral character of the participants. A not untypical proponent of the latter view, arguing against organizational changes in the market for psychiatric residencies, says, "Hopefully, the profession of medicine and its specialty of psychiatry will seek to avoid the mud of the marketplace" (Myers 1979, p. 110).

[32] However, it is not premature to speculate fairly specifically on these features. Recall that salary is not the dominant factor in determining students' preferences for hospital programs. Since students are making a choice that has implications for their entire subsequent career as physicians, it is unsurprising that factors such as the reputation of the hospital as a training facility should play a critical role in determining their preferences. For a hospital to raise its reputation significantly would typically require expenses far beyond those associated with the salaries of residents. For many hospitals it must therefore be more feasible to fill positions that might otherwise go to residents with more senior physicians, for whom the hospital's educational role is not a concern

Darrach, William. "Letter." *Bull. Assoc. American Medical Colleges* 2 (January 1927): 68.

Darley, Ward. "The Seventh Annual Report of the National Intern Matching Program." *J. Medical Educ* 34 (January 1959): 38–46.

Demange, Gabrielle, and Gale, David. "A Strategy-proof Allocation Mechanism for Two-sided Matching Markets." Mimeographed. Paper presented at the IMSSS-Economics workshop Stanford, Calif.: Stanford Univ , 1983.

Dubins, L. E., and Freedman, D. A. "Machiavelli and the Gale-Shapley Algorithm." *American Math. Monthly* 88 (August/September 1981): 485–94

Elias, Anthony David, and Elias, Ellen Roy. "Matching of Couples in the NRMP." *New England J. Medicine* 302 (June 19, 1980): 1425–26

Fitz, Reginald. "Concerning Interns and Internships." *J. Assoc. American Medical Colleges* 14 (March 1939): 101–4.

Gale, David, and Shapley, Lloyd S. "College Admissions and the Stability of Marriage." *American Math. Monthly* 69 (January 1962) 9–15

Gale, David, and Sotomayor, M "Some Remarks on the Stable Matching Problem." Mimeographed. 1983.

Gillies, D. B. "Locations of Solutions." In *Report of an Informal Conference on the Theory of* N*-Person Games*, edited by Harold W. Kuhn Princeton, N J · Princeton Univ., 1953. (*a*)

———. "Some Theorems on *N*-Person Games." Ph.D dissertation. Princeton Univ., 1953. (*b*)

Graettinger, John S. "Graduate Medical Education Viewed from the National Intern and Resident Matching Program " *J Medical Educ* 51 (September 1976): 703–15.

Graettinger, John S., and Peranson, Elliott "The Matching Program " *New England J. Medicine* 304 (May 7, 1981)· 1163–65

Green, Jerry R., and Laffont, Jean-Jacques. *Incentives in Public Decision-Making* Amsterdam: North-Holland 1979

Hylland, Aanund, and Zeckhauser R.

Individuals to Positions " *J P E*

"The Internship Matching Plan

"Internships." *J. Assoc. American M*

Kelso, Alexander S., Jr., and Craw

Formation, and Gross Substit

1483–1504.

Leonard, Herman B. "Elicitation

of Individuals to Positions " *J P*

"Modification of Cooperative Plan for

American Medical Colleges 25, no. 6, pt 1

Mullin, F. J. "A Proposal for Supplementing the Cooperative Plan for Appointment of Interns." *J. Assoc American Medical Colleges* 25, no 6, pt 1 (November 1950): 437–42.

Mullin, F. J., and Stalnaker, John M. "The Matching Plan for Internship Appointment." *J Medical Educ* 26 (September 1951) 341–45.

———. "The Matching Plan for Internship Placement. A Report of the First Year's Experience." *J. Medical Educ* 27 (May 1952)· 193–200.

Myers, J. Martin. "Reply." *J. Clinical Psychiatry* 40 (February 1979)· 110.

NIRMP Directory. Evanston, Ill : National Resident Matching Program, 1979

NRMP Directory. Evanston, Ill · National Resident Matching Program, 1983

Quinzii, Martine. "Core and Competitive Equilibria with Indivisibilities " Mimeographed. Paris: Ecole Polytechnique, Laboratoire d Econometrie, 1982.

"Report of the Committee on Internships of the Association of American Medical Colleges." *J. Assoc. American Medical Colleges* 21 (May 1946): 178.

Ritz, Zvi. "Incentives and Stability in Some Two-sided Economic and Social Models." Mimeographed. Urbana-Champaign: Univ. Illinois, Dept. Bus. Admin., 1982.

Roth, Alvin E. "The Economics of Matching: Stability and Incentives." *Math. Operations Res.* 7 (November 1982): 617–28. (*a*)

———. "Incentive Compatibility in a Market with Indivisible Goods." *Econ. Letters* 9, no. 2 (1982): 127–32. (*b*)

———. "Conflict and Coincidence of Interest in Job Matching: Some New Results and Open Questions." *Math. Operations Res.* (1984), in press. (*a*)

———. "Incentives in the College Admissions Problem and Related Two-sided Markets." Working Paper no. 173. Univ. Pittsburgh, Dept. Econ., 1984. (*b*)

———. "Misrepresentation and Stability in the Marriage Problem." *J. Econ. Theory* (1984), in press. (*c*)

———. "Stability and Polarization of Interests in Job Matching." *Econometrica* 52 (January 1984): 47–57. (*d*)

Roth, Alvin E., and Postlewaite, Andrew. "Weak versus Strong Domination in a Market with Indivisible Goods." *J. Math. Econ.* 4 (August 1977): 131–37.

Rutkow, Ira M., and Glasgow, Arthur H. "How Medical Students View the Application and Reviewing Procedure for Surgical Residency." *J. Medical Educ.* 53 (June 1978): 505–7.

Shapley, Lloyd S. "Open Questions." In *Report of an Informal Conference on the Theory of N-Person Games*, edited by Harold W. Kuhn. Princeton, N.J.: Princeton Univ., 1953.

Shapley, Lloyd S., and Scarf, Herbert. "On Cores and Indivisibility." *J. Math. Econ.* 1 (March 1974): 23–28.

Stalnaker, John M. "The Matching Program for Intern Placement: The Second Year of Operation." *J. Medical Educ.* 28 (November 1953): 13–19.

Sudarshan, Alex, and Zisook, Sidney. "National Resident Matching Program." *New England J. Medicine* 305 (August 27, 1981): 525–26.

Turner, Joseph. "Intern Selection: Wanted, an Orderly Plan." *J. Assoc. American Medical Colleges* 20 (January 1945): 26–32.

"Uniform Intern Placement Plan for 1949." *J. Assoc. American Medical Colleges* 24 (January 1949): 42.

Williams, Kevin Jon; Werth, Victoria P.; and Wolff, John A. "An Analysis of the Resident Match." *New England J. Medicine* 304 (May 7, 1981): 1165–66.

On Interindustry Differences in Absolute Productivity

William J. Baumol

Princeton University and New York University

Edward N. Wolff

New York University

In the literature, comparisons between absolute levels of productivity in different industries occur frequently. This paper explores the meaning and possible significance of such measures. Prices must be used to permit the required comparison of the outputs of different industries. It is shown that base-year weights produce a valid measure of productivity growth but not of absolute productivity level. Current price weights do yield a valid index of absolute productivity but one that tends to be equal for all industries because of the general equilibrium mechanism that
productivity to high-productivity

Comparisons of absolute levels o
To cite just a few examples, Nord
base-year sectoral price deflators,
ity in public utilities and the finance, ins
was over three times the mean productivity level in the economy in
1971, while that in services and the government sector was about half
Denison (1973), using an overall price deflator, calculated that the

We are most grateful for their very valuable comments to David Dollar, John R Norsworthy, Albert Rees, and Mark Shankerman. We also appreciate deeply the support of the Division of Information Science and Technology of the National Science Foundation and the C. V. Starr Center for Applied Economics, New York University

[*Journal of Political Economy*, 1984, vol 92, no 6]
© 1984 by The University of Chicago All rights reserved 0022-3808 84 9206 0003$01 50

level of productivity in sectors producing commodities was about equal to that of the service sector in 1929 and was about 10 percent higher in 1969. Thurow (1980), using base-year sectoral deflators, found that in 1972 output per man-hour was highest in the finance, insurance, and real estate, public utilities, and communications sectors (about $20.00 per hour in 1972 dollars) and lowest in retail trade and services (about $5.00 per hour). Kutscher, Mark, and Norsworthy (1977), Denison (1979*b*), and Baily (1982) all use differences in the level of productivity between sectors, based on base-year sectoral deflators, in order to calculate the effect of shifts in employment on overall productivity growth.

Such analyses raise the issue of the meaning of such differences in absolute levels of sectoral productivity. Is there any meaning at all to comparison of number of apples produced per hour with number of haircuts per hour? On the surface it appears that apples can be compared with haircuts in terms of the *value* of apples produced per hour with the value of haircuts produced per hour. Yet, in equilibrium, under conditions of perfect competition, as we will see, the average productivity of all inputs taken together (i.e., total factor productivity), if measured in terms of current values, must of necessity be equal among industries. Thus even a comparison of absolute productivities in terms of output values would appear to be meaningless.

Is there, then, any defensible way to interpret the notion of absolute industry productivity and to define its magnitude? This is the subject of the present paper. We will focus on two candidate measures of absolute productivity, both using values of outputs for their interindustry comparisons. The first, the *base-year* measure, uses the prices prevailing in some base-year period, keeping those price weights constant in the index calculations for all other years. The second candidate measure of absolute productivity, the deflated index, instead uses current prices deflated by some index of the economy's overall price level.

We will show that the base-year measure, while very defensible as an indicator of rate of productivity *growth*, is a complete failure for our purposes: it yields a figure for absolute level of productivity for an industry that is totally arbitrary and that we can make as large or as small as we like. Thus, consider two industries, a and b, and suppose that a's productivity as measured by the base-year index is, say, 3 percent larger than b's, year in and year out. Then by appropriate choice of base year it is possible to conclude that a's *absolute* productivity level today is the same as b's or is five times as large as b's or is 3.25 times as great as b's. Or we can instead decide in advance that we would like a's absolute productivity level to turn out to be 2.8 times as large as b's and then determine the base year that gives us this ratio. In sum, the base-year index, despite its great value as an indicator of productivity growth, is totally useless as a comparative measure of the absolute productivities of two sectors of the economy.

We will see, in contrast, that although the deflated index of absolute productivity tends to assign the same absolute productivity figures to all economic sectors, and certainly does so if those sectors are perfectly competitive, that measure actually reports a substantive piece of economic information—the marginal welfare productivity of each sector of the economy. For what the general equilibrium mechanism of perfect competition does is to redistribute inputs and consumer purchases among the economy's different products in such a way that at the margin the absolute welfare productivity of any input is the same in all industries. Moreover, that productivity figure will grow in each sector at the same rate as an index of the real compensation of the pertinent inputs.

We will conclude, then, that a deflated index of absolute productivity does have substantial content, though a comparison of the figure for different industries can tell us little, since market forces make that index behave similarly for all industries.

On the other hand, since the base-year index can tell us nothing about an industry's absolute productivity, it is useless for interindustry comparisons. We are left, then, with the conclusion that there is a way of defining an industry's absolute productivity that does convey substantive information, although we have not been able to find a useful measure to distinguish the abs
of one industry from that of anoth
ture cited here earlier, therefore,
tion.

In discussing measures of prod
more obvious types of indices used
productivity as the value of the per
divided by the value of the vector of inp
avoid sophisticated constructs, such as the Divisia index, with their strong theoretical foundations and their valuable analytical properties (on this material, see, e.g., Diewert [1980]; Gollop and Jorgenson [1980]). Accordingly, in much of what we discuss it will be useful to view our main productivity indices not as measures of the shift in the production frontier that occurs with the passage of time, but rather as indicators of the improvement in consumer and producer welfare derivable from a given collection of inputs. This is the ultimate justification for our attention to the current market values of outputs and inputs.

I. Absolute Productivity, Aggregation, and the Use of Price Weights

Absolute productivity may be defined, roughly, as some measure of output per unit of input.[1] One problem raised by this definition is that probably no firm and certainly no industry or sector of the economy, such as "the services," produces only a single output or uses only one input. Moreover, since our main concern is interindustry comparison, we are faced with the issue of achieving comparable output measures when each industry's output composition is very different from every other's. There is, of course, no one correct or "best" way of dealing with this problem. However, the obvious approach, which is rather widely used and which, as we will confirm, makes considerable sense, is to use prices as weights for the aggregation of outputs and for the aggregation of inputs.

Legitimate price weights for productivity measurement obviously must avoid the distorting influence of inflation. There are (at least) two ways in which this can be done, and each of them turns out to yield a measure of productivity whose interpretation is significant but very different from that of the other: (1) The base-year approach uses as its weights for output i, p_{i0}, the price of product i in some base period, 0, instead of p_{it}, the nominal price of i in period t. (2) The deflated *price* approach replaces p_{it} by p_{it}/p_t, where p_t is some sort of overall price index such as $\Sigma\, p_{it}y_{it}/\Sigma\, p_{i0}y_{it}$. Here, since we are seeking some sort of measure of real value of output to consumers as a class, it is appropriate to take as p_t some index of the overall price level of the economy, for example, the consumer price index, rather than any measure of the behavior of prices in the sector of the economy where the outputs in question are produced.[2] A similar pair of options, namely, w_{k0} and w_{kt}/w_t, is, of course, available for the derivation of a real value corresponding to the nominal price, w_{kt}, of input k in period t.

Before going on to our analysis proper, it is helpful, simply as a reference point, to derive a relationship for the spurious measure of productivity that would result if nominal prices were used to aggregate outputs and inputs. We obtain the nominal total factor productivity of sector s:

---

[1] This way of putting the matter inevitably must cause some problems for those who will accept as an increase in productivity only a shift in the production set, i.e., an inward shift in the production isoquants in input space. If absolute productivity is defined as in the text, it can be increased by such phenomena as scale economies or elimination of X inefficiencies that move the economy toward the production frontier. The point in the Divisia index approach is that for wide classes of technological change it does isolate the shifts in the production isoquants and eliminates changes in the output-input ratio attributable to developments that do not affect the production function.

[2] As we will see presently, the use of a production-sector price index for the deflation process gives us results very similar to those stemming from base-year prices.

$$N_{fst} = \frac{\sum_{i \in S} p_{it} y_{it}}{\sum_{k \in S} w_{kt} x_{kt}}, \tag{1}$$

where S is the set of outputs produced and inputs used in sector s, y_{it} is the quantity of the ith good produced by s in period t, and x_{kt} is the quantity of input k it uses.

To bring out our general equilibrium results most sharply, in most of the remainder of the analysis we will work on the assumption that the economy in each period achieves a perfectly competitive equilibrium, though that equilibrium, of course, changes from period to period.

A moment's consideration shows that N_{fst}, the ratio of value of output to value of input, is a measure of the economic profitability of sector s. Hence, since profits are always zero in competitive equilibrium, we must have

$$N_{fst} = 1 \quad \text{in every sector } s. \tag{2}$$

This relationship will prove helpful in our later discussions.

II. Base-Year Productivity Measures in Literature

The most obvious way to go abou
of inflation is to use base-year pri
is, to keep the price weights unch
yields the obvious formula:

$$B_{fst} = \frac{\sum_{i \in S} p_{i0}}{\sum_{k \in S} w_{k0} x_{kt}}.$$

Before subjecting this index to a formal analysis, let us first show that it is indeed something that has been used by economists who work in the area.

Because explicit statements specifying the productivity formulae used in the empirical literature are surprisingly infrequent, it is not always easy to judge which authors have used equation (3) or something close to it as the basis for some or all of their productivity calculations. Yet, it seems clear that (3) or a near relative is frequently employed for the purpose. At least five studies have used such expressions: Stigler (1947), Nordhaus (1972), Kutscher et al. (1977), Thurow (1980), and Baily (1982). Actually, at least some of these authors use an approach that at first appears to be a compromise between our base period and our deflation method—rather than employing base-period prices, they employ current prices deflated by an index of the price level for just the sector with which their calculation is concerned. But the resulting behavior of their productivity index is essentially that of a base-period price. In particular, if they employ the price index

$$p_{st} = \frac{\sum_{i\in S} p_{it}y_{it}}{\sum_{i\in S} p_{i0}y_{it}}, \tag{4}$$

then the value of the sector's output that they obtain becomes

$$\sum_{i\in S}\left(\frac{p_{it}}{p_{st}}\right)y_{it} = \sum_{i\in S}\frac{p_{it}y_{it}}{p_{st}} \equiv \sum_{i\in S} p_{i0}y_{it}, \tag{5}$$

which is, of course, exactly the base-year measure of aggregate output. Clearly, any other price index whose behavior approximates that of B_{st} will yield roughly similar results. Intuitively, the point is that deflation by a sectoral price index returns the "average" price of that sector to its base-year value, and that is why deflation by a sectoral price index gives roughly the same results as use of base-year prices as weights.[3]

III. The Base-Year Measure and Absolute Sectoral Productivity

Let us turn now to the workings of the base-year productivity measure (3). We will argue that though it is, of course, by no means the only measure that can qualify for the task, this index, B_{jst}, is a defensible measure of the growth of physical productivity in any sector of the economy.[4] That is, it can be interpreted as a reasonable indicator of the growth in the sheer quantity of goods and services that can be turned out by that sector with a given input bundle. However, we will suggest that there are two things this index cannot legitimately do: (i) It is not an acceptable indicator of growth in *welfare* productivity, that is, in the ability of a given bundle of inputs to contribute to the welfare of the community (roughly, to consumers' and producers' surpluses) where inputs themselves are measured in welfare terms, that is, in terms of their social opportunity cost. For evaluation of growth in welfare productivity we will have to turn to another index that will be described presently. (ii) Moreover, though the base-year index is a measure of physical productivity growth, it is totally useless as a measure of absolute productivity level in any sector or of the difference in the absolute productivity levels of any two sectors.

---

[3] All the authors we are discussing seem to use a procedure that also differs from (3) in another respect. They work with gross product originating (GPO) by sector as their measure of sectoral output. Their productivity measure is, essentially, sectoral output deflated by the sectoral price deflator less intermediate inputs deflated by the corresponding input price deflators. Thus, value in each product line is evaluated at its own uninflated prices. The method is often referred to as "double deflation," since the outputs and intermediate goods are deflated separately. It is also sometimes called "value added in constant prices." This measure, which we may symbolize by v_{jst}, has the same properties we will show for B_{jst}. In particular, if its rate of growth, r, is constant and the economy is always in competitive equilibrium, $v_{js0} = 1, \forall s$, and $v_{jst} = v_{js0}e^{rt} = e^{rt}$

To see in what sense B_{fst} serves as a measure of physical output growth, let us subject it to the obvious test. Let us observe its performance in a case where physical productivity growth is easily definable and measurable—that in which all input quantities grow in some fixed proportion, e^{qt}, and outputs all grow in some other common proportion, e^{rt}, over some period t. Then, obviously, by (3),

$$B_{fst} = \frac{\sum e^{rt}p_{i0}y_{i0}}{\sum e^{qt}w_{k0}x_{k0}} = \frac{e^{rt}}{e^{qt}}B_{fs0} = e^{(r-q)t}B_{fs0}, \tag{6}$$

so that B_{fst} will have grown at the rate $r - q$, just as physical productivity has. Analogously, when outputs and/or inputs do not grow proportionately, B_{fst} can clearly be interpreted as the growth of a weighted average of product per unit bundle of input, with all of the usual reservations about the aggreg

index number.

It must be reiterated that B_{fst} can

production frontier of sector s, but

quences of such a change with the

in input or output proportions,

---

[4] Here, we must reemphasize that we are not

index in measuring growth in physical productivity

former can, reasonably, be interpreted as one possible way of calculating the

in physical output obtainable from a fixed bundle of inputs

[5] Indeed, one cannot normally measure the rate of shift of the production frontier statistically without (quite legitimately) adopting some assumption that removes a good part of the problem. Thus, if one assumes that the production relationship is of the form $y = e^{vt}f(\mathbf{x}_t)$, where $\mathbf{x}_t$ is the vector of inputs, absolute productivity is measured implicitly by $\pi = y/\bar{x}_t$, where $\bar{x}_t$ is some index of input usage and productivity growth equals $\dot{\pi}/\pi = v + d\ln[f(\mathbf{x}_t)/\bar{x}_t]/dt \cong v$ if $f(\,)$ is approximately linearly homogeneous. Thus, v is the correct measure of the change in the rate of growth of productivity measured as a shift in the production frontier under the preceding assumptions (which are similar to those widely employed in the literature). However, the assumed mathematical form of the production relationship clearly restricts very substantially the nature of the technological change that can be experienced by the economy of this model

Thus, the physical productivity growth measured by B_{fst} is to be interpreted quite differently from that corresponding purely to a shift in the production frontier. Both types of physical productivity growth concepts, B_{fst} and a measure of the shift in the frontier, have legitimate interpretations, but they are not the same.

Next, we note very briefly why B_{fst} is not a defensible measure of growth in welfare productivity. The point is that in period t the price vector $\mathbf{p}_0$ is a vector of obsolete prices and hence represents consumer evaluations of the products that are no longer relevant if tastes have changed. A commodity i whose price has since fallen sharply, because preferences have shifted to other goods, is still evaluated by B_{fst} at its high earlier price, p_{i0}, as though its marginal welfare contribution were still as great as it once was. This suggests that current prices are a better way of going about measuring growth in welfare productivity, and that is indeed the approach we will employ for the purpose.

Finally, and most important for our main subject, let us see now why B_{fst} is not an acceptable measure either of *absolute* physical productivity of any sector of the economy (even if it is a legitimate measure of its growth) or of the relative productivities (i.e., the ratio of the absolute productivities) of any two sectors. First we note that by (1) $B_{fs0} \equiv N_{fs0}$; that is, in the base year there is, by definition, no difference between the base year and the nominal productivity figures. Thus, in competitive equilibrium, by (1) and (2),

$$B_{fs0} \equiv N_{fs0} = 1. \tag{7}$$

Next, consider two sectors of the economy, a and b, with respective constant productivity growth rates u and v, and take $u \geqslant v$. Then clearly, by (6) and (7),

$$B_{fat} = e^{ut}, \quad B_{fbt} = e^{vt}, \quad \frac{B_{fat}}{B_{fbt}} = e^{(u-v)t}. \tag{8}$$

These, then, are the absolute and relative productivity figures for the two sectors implicit in the B_{fst} measure. But a glance at (8) reveals their fatal flaw. For they are *totally* dependent on the choice of base year in a manner far more serious than that normally entailed by the arbitrary choice of a base period for some index number. The magnitude of an index number of GNP, for example, depends heavily, as we know, on the state of the economy at the particular date selected as the base year, and it may make current GNP figures seem particularly high or low by comparison. In contrast, the values of the expressions in (8) depend not on economic conditions in the base period, but only on t, the amount of time since elapsed. Thus, the three numbers that emerge from (8) are simply an artifact of the number of periods earlier that have been chosen for the base. We can arbitrarily make these numbers as large as we want simply by taking the base year to lie sufficiently far in the past. Similarly, we can make it as small as we like by selecting a base year sufficiently recent, thus making t correspondingly small. Thus we can make the current ratio of absolute productivities in sectors a and b equal to any arbitrary number, $k \geqslant 1$, simply by choosing the base period to be the date t^* periods earlier, where $t^* = [\ln(k)]/(u - v)$, so that by (8) $B_{fat^*}/B_{fbt^*} = e^{(u-v)t^*} \equiv k$. In sum, even though (8) clearly gives us the percentage growth rates of physical productivity, u and v, of the two sectors, the absolute and relative productivity figures that emerge are completely arbitrary and devoid of economic content.[6]

IV. The Deflated Productivity Index

Economic analysis has long emphasized that sheer physical output—the economy's miscellaneous outpouring of objects and services—is not its basic concern. Rather, what is ultimately pertinent is the economy's ability to serve the desires of individuals in their roles of consumers and producers. If an economy were to double the amount of commodity X it can produce per hour but consumers' tastes had meanwhile changed and good X was valued far less than before, all other things being equal, productivity in this industry *may in a sense* have fallen. For, if productivity is to serve as a measure of an economy's ability to support a high standard of living the measure
employed for the purpose must in [illegible]
contribution of the outputs rather [illegible]

One obvious way to approach th [illegible]
mation to the desired standard the [illegible]
of production, meaning that one e [illegible]
one another, at the ratio of their m [illegible]
at their relative marginal opportunit [illegible]
made operational, at least in our world of [illegible]
librium. For there, as the textbooks tell us, marginal utilities must [illegible]

---

[6] It should be clear by the same reasoning why, as indicated in an earlier note, similar problems arise if instead of base-year *prices we deflate prices by sectoral price indices* and interpret the resulting figures as indices of absolute or relative sectoral productivity. It may also be noted that, in terms of our subject, the most interesting reason for *which prices vary over time is technical change. As a result, the long-run competitive* equilibrium position of the economy keeps shifting. Equation (7) thus refers to one such equilibrium position with zero profits in all industries

[7] *In this respect, a productivity index that measures the shift in the production* frontier has things both ways—for such a shift represents an ability to increase both physical output *and* welfare (assuming that consumers are not sated in everything) An outward shift in the entire length of the frontier means that the economy has increased at least its *capacity* to enhance welfare This does not necessarily follow, of course, if only part of the frontier shifts outward.

proportionate to real output prices and marginal opportunity costs of inputs must be proportionate to real input prices. This immediately leads us to our second productivity measure, the deflated index of total factor productivity:[8]

$$D_{fst} = \frac{\sum_{i \in s} (p_{it}/p_t) y_{it}}{\sum_{k \in s} (w_{kt}/w_t) x_{kt}}, \tag{9}$$

where p_t and w_t are, respectively, any of the standard indices of the economy's overall level of output and input prices. As already noted, here we employ general price levels for the entire economy rather than those of any particular sector. For we are concerned with the evaluation of the economy's body of consumers, as a group, of the outputs of sector s, and the opportunity cost to the economy as a whole of the bundle of inputs used by sector s. The use of the marginal measure of welfare contribution does have a conceptual limitation, as the pertinent indicator of welfare, which can be quite serious,[9] but it is at least a step in the right direction.

We may note that it is more difficult to find in the empirical literature examples of the use of the D_{fst} measure or approximations to it than in the case of B_{fst}. Yet something like it seems to be implicit in the discussions of what has been called "the resource reallocation effect" (see, e.g., Denison 1979*a*, 1979*b*, 1982; Norsworthy, Harper, and Kunze 1979; Gollop 1982). Here, overall productivity is taken to be increased by a more efficient allocation of resources in a shift toward equilibrium. The most straightforward and simple use of the approach appears in Denison's writings.

V. Behavior of the Deflated Index

We will show next that when D_{fst} is used as the productivity index, the value of absolute productivity has an inherent tendency toward similarity, from sector to sector, and for the different sectors to maintain a rough parity with one another. This curious behavior is a consequence of the use of current values in measuring productivity. To see the connection suppose, for example, that the physical productivity of labor in clockmaking doubles. If, as a result, the price of a clock falls by half, the *value* of the output of a worker (his value or nominal productivity) will remain unchanged. This is the normal qualitative relationship, since whenever there is a rise in the physical productivity of the inputs used to produce an item, the price of the item will tend to fall.

---

[8] The logic of the analysis is unaffected if the index is modified to relate to the value added rather than the gross output of sector s. All that is required is a corresponding change in the interpretation of the y_{it} and the x_{kt}.

[9] The point is that D_{fst} is a legitimate measure of growth in *welfare* productivity only for marginal changes. This is so because price is a measure of consumers' benefits only at the margin, since larger changes will also affect consumers' surplus. A similar comment applies on the input side. For larger changes the measurement of growth in welfare productivity is, consequently, far more complex.

Whether the deflated productivity of the inputs employed in making that item rises or falls may, then, appear to depend on whether its price falls faster or more slowly than the real relative productivity of its inputs rises, but matters are not so indeterminate. Competition and capital flows will tend to equalize the ratios of the value of output to the value of inputs for all sectors, making for equality in the value productivities of their inputs. For this reason, intersectoral comparisons of absolute productivity should tend to show that the differences are small and do not change greatly with the passage of time, even though physical productivity in some grows far more rapidly than in others.

The point in its strongest form is a model of moving competitive equilibrium. A comparison of (9) and (1) indicates at once that

$$D_{fst} \equiv N_{fst}\left(\frac{w_t}{p_t}\right), \tag{10}$$

for the deflated index simply is the nominal index with output prices divided by output price index p_t and in[illegible] number w_t. However, in perfectly [illegible] (2), $N_{fst} = 1$, so that, from (10),

$$D_{fst} [illegible]$$

Because the value of the ratio [illegible] time, it is clear that D_{fst}, unlike N_f [illegible] time. However, since this ratio does not vary [illegible] from industry to industry, a deflated (real) measure of productivity level implies a tendency toward equality of the total factor productivity figures for the different sectors, with absolute equality always holding in competitive equilibrium.

It will still be true, even if paradoxical, that at least in a perfectly competitive economy, even if physical productivity in one sector grows persistently faster than that of another, real productivity in the two sectors as measured by D_{fst} will begin at the same level and move together in lockstep through time. Use of this index, for example, may indicate that productivity in the service sector is equal to that in goods production and remains equal over time.

It should be noted that D_{fst} must at least be accepted as a reasonably satisfactory indicator of total factor productivity growth in the economy as a whole, since it amounts to an index of real compensation of inputs. For its value, w_t/p_t, is simply average input price over average output price.

Before we explore the source of the behavior of D_{fst}, it is worth noting an implication of this uniformity of D_{fst} in all sectors and the diversity of values of B_{fst} for empirical evaluation of the effect on overall productivity growth of shifts in the economy's labor and other inputs among the sectors. Our analysis suggests that if current prices (real or nominal) are used as weights in a statistical calculation, then intersectoral shifts in input uses will be estimated to have had little effect on productivity. This is so because the resources will have moved from a sector s with the deflated productivity level D_{fst} (or N_{fst}) to another sector, s', with about the same deflated productivity level, and so the calculated net effect is virtually certain to be small. Yet this result may fail to capture the full physical effect of the resource shift, because if the shift moves inputs toward sectors with high growth rates in physical productivity, it will increase w_t/p_t and hence the value of D_{fst} in each and every sector s. On the other hand, if base-year prices are used, depending on the choice of the base period, then B_{fst} and $B_{fs't}$ are likely to differ substantially. Thus if, say, $B_{fst} - B_{fs't}$ is large, then a shift of resources from s to s' is likely to affect the *calculated* productivity growth rate substantially, by amounts that differ among studies because of differences in the value of t.

The results of the available empirical studies are roughly consistent with these observations, though the current price formulas used do not quite correspond to D_{fst} or N_{fst}. Base-year prices are used by Nordhaus (1972), Kutscher et al. (1977), Thurow (1980), and Baily (1982), who conclude that input shifts account, respectively, for 79, 23, 45–50, and zero percent of the slowdown in overall productivity growth for the (different) periods they consider. On the other hand, in a portion of their formulas, current price weights are used by Nordhaus (in another part of his 1972 study) and by Gollop (1982), each of whom then obtains an estimated effect approximately equal to zero.

VI. The Paradox Resolved: The Market as the Distributor of Welfare Productivity Growth

One may well be tempted to conclude that the discounted index of absolute sectoral productivity is as devoid of content as the base-year measure. What sense can there be to a sectoral measure of total factor productivity that inevitably maintains parity among sectors through time when the economy remains in a moving competitive equilibrium? Does this not indicate that D_{fst} is an unacceptable measure of absolute or intersectoral differences in productivity—that we should be looking for something else?

We will show next that, on the contrary, D_{fst} has a very sensible and straightforward economic interpretation that emerges from elementary economics. That interpretation also tells us more about the role of the base-year index. The picture that emerges is, in brief, the following: In competitive equilibrium the deflated index, D_{fst}, for each sector grows at the same rate as an average of the rates of physical productivity growths of the economy. Moreover, the absolute value of each D_{fst} is equal to the real earnings of an average input in the economy (the real wage in a one-input model). In sum, the market mechanism readjusts the prices of the different products with their different growth rates of physical productivity so as to shift inputs and quantities consumed in such a way that the growth in marginal welfare yields of all the inputs is equalized. In effect, then, the competitive mechanism translates the physical growth achievements of the economy into increases in welfare contributions of inputs and, in the process, equalizes them.

Thus the tendency toward equality of the D_{fst} does *not* make nonsense of them. They are legitimately interpretable as measures of the economy's mean standard of real compensation of an input, that is, of its average living standard. Moreover, D_{fst} is the "right" measure of the sector's economic productivity [illegible] miscellaneous physical objects [illegible] consumers desire them. Rather [illegible] product as it is currently evalua[illegible] In short, in economic terms, the [illegible] just about the only measure that [illegible] absolute productivity. If it does [illegible] ences in absolute productivities among [illegible] those differences are simply not there.

Thus, as will now be shown in somewhat more detail, the equality of D_{fst} in all economic sectors is no quirk of the measure, but an accurate description of an underlying behavioral phenomenon. The explanation, as already indicated, lies in the workings of the invisible hand, which adjusts prices and reallocates input usage and consumption patterns to adapt them to changes in the pertinent relative productivities. Price reactions have already been discussed, and the behavior of input usage is also straightforward. Every elementary textbook reminds us that the competitive mechanism tends to reallocate inputs whenever such a transfer increases their marginal value products. Thus, suppose the marginal value productivities of labor in sectors s and s' were equal in period t but that in s rises in the next period while that in s' remains constant. We then expect a shift of the labor force from s' to s up to the point at which its marginal revenue product in the two sectors is equalized once again.

Of course, this does not mean that the pattern of physical productivity growth will control these movements by itself, because as we have noted, relative price in the sector whose physical productivity lags will rise by an amount corresponding to its rising relative marginal cost, and it is the marginal value products, not marginal physical products, that are equated by the reallocation of inputs. Yet it remains true that input usage will be adjusted by market forces to intersectoral disparities in productivity growth, and that at the margin it will tend to iron out these disparities.

But how can the relative welfare contribution of the inputs of a sector that is lagging in physical productivity be adjusted to rise commensurately with that in sectors whose productivity rises rapidly? The answer is that the equilibration process shifts not only inputs, but also consumers, among the different output sectors. Rising relative prices in the more stagnant sector drive away those consumers who derive relatively little benefit from its products, leaving behind only those to whom their value is relatively high. For example, the increasing relative costs and prices that have characterized the performing arts throughout much of their history have presumably driven out of the audiences those who previously were on the borderline between attendance and other activities, leaving an audience composed increasingly of hard-core lovers of music, drama, or dance. As prices rise, the pleasure derived from attendance by a borderline potential ticket purchaser must therefore increase commensurately, and so the marginal welfare yield of an attendee's presence will rise in step with price.

The intersectoral equality of the D_{fst}, the deflated total factor productivities, then, is nothing but a reflection of the efficiency properties of general equilibrium under perfect competition. Inequality of D_{fst} and $D_{fs't}$ represents an inefficiency that can be remedied by a shift in resources and consumers between the two sectors.

VII. Physical Productivity Growth as the Source of Welfare Productivity Growth

In this way, the competitive mechanism serves as an instrument that redistributes physical productivity growth, transforming it into value productivity growth distributed equally among all sectors. Even if "physical" productivity in the services[10] had grown, not just more slowly than that in manufacturing, but not at all, under perfect competition the equilibrating process would have seen to it that both grew in value at the rate of increase in real wages, w_t/p_t, as (11) tells us. The growth of physical productivity in manufacturing becomes a sort of fund in which, in equilibrium, both manufacturing and the services share equally. That is the meaning of (11) and the conclusion that value productivity will proceed at the same pace in all sectors.

While increased welfare is, from our point of view, the objective of productivity growth, we will show next in what sense its source is the (uneven) expansion in physical productivity that can be evaluated by the base-year measure, B_{fst}. This is all expressed formally with the aid of a straightforward proposition. It is easy to prove that if we select as the price indices used in defining D_{fst} and D_{ft} the Paasche index numbers using current outputs and inputs as weights (and again ignore the traditional problems of index numbers, particularly if tastes change), then for the economy as a whole the index of physical productivity, B_{ft}, and the index of discounted productivity, D_{ft}, must be identical. That is, we must have

$$B_{ft} \equiv D_{ft}. \tag{12}$$

Here

$$B_{ft} \equiv \frac{\sum\limits_{\forall i} p_{i0} y_{it}}{\sum\limits_{\forall i} w_{k0} x_{kt}} \equiv \sum_{s} \frac{\sum\limits_{k \in s} w_{k0} x_{kt}}{\sum\limits_{\forall k}}$$

where θ_{st} = sector s's share of

year prices, so that B_{ft} is an av

Similarly (under competitive eq

$$D_{ft} \equiv \frac{\sum\limits_{\forall i} (p_{it}/p_t) y_{it}}{\sum\limits_{\forall k} (w_{kt}/w_t) x_{kt}} = \frac{w_t}{p_t} = D_{fst}, \ \forall s \tag{12b}$$

We obtain directly, from the definition of the Paasche index, summing over all i, k,

$$D_{ft} \equiv \Sigma\, p_{it} y_{it} \frac{\Sigma\, p_{i0} y_{it}}{\Sigma\, p_{it} y_{it}} \Big/ \Sigma\, w_{kt} x_{kt} \frac{\Sigma\, w_{k0} x_{kt}}{\Sigma\, w_{kt} x_{kt}} \equiv \frac{\Sigma\, p_{i0} y_{it}}{\Sigma\, w_{k0} x_{kt}} \equiv B_{ft}.$$

---

[10] The concept of "physical productivity" in the services is, of course, not to be taken literally. We mean by it such things as number of books circulated per librarian-hour or number of persons in the audience per musician-hour.

as equation (12) asserts. It follows from (12) that for the economy as a whole the two productivity figures must grow at the same percentage rate. Thus, if the economy's overall physical productivity were constant ($d \ln B_{ft} = 0$), then in competitive equilibrium the growth of deflated productivity, $D_{fst} = w_t/p_t$, must be zero in every sector, s. Similarly, when $d \ln B_{ft} \neq 0$, in competitive equilibrium each D_{fst} must increase roughly as the *average* B_{fst} does (unless this effect is offset by productivity-changing shifts in the allocation of inputs), as the Appendix shows. Thus, (12) is basically the formal justification of the parable in which B_{fst} is used to measure the source of growth in welfare productivity, while D_{fst} indicates the way in which it is redistributed and diffused through the economy by the market mechanism.

VIII. Concluding Comment

Our search for a measure that permits comparison of the absolute productivities of different sectors of the economy has, we believe, been illuminating. But in one respect it has clearly been a resounding failure. For the one thing we have failed to provide is a measure on whose basis it is possible to say convincingly that sector a is 1.5 times as productive as sector b. In retrospect we are convinced that the difficulty is inherent in the nature of the issue—which requires us to compare the apples and oranges that are the produce of the different sectors. As we know, in economics a significant comparison of apples and oranges *is* possible, but only in terms of the values of those items. There, however, is the rub, because we have seen that once values are used for the purpose, we are driven to the conclusion that absolute productivities in different sectors have an inherent tendency to equality. This conclusion, it has been shown, is economically defensible. But it cannot be very satisfying to those who wish to make absolute intersectoral comparisons with the aid of input and output data.

Appendix

Overall Productivity Growth as Average of Sectoral Physical Productivity Growth

It was asserted in the text that the market mechanism serves to distribute equally, in value terms, the differing physical growth rates of the various sectors. In this Appendix this relationship is derived explicitly,[11] using current-year prices in the price and wage indices p_t and w_t employed to measure the economy's productivity in (9). In other words, we use a Paasche index of real wages as our deflated index of productivity, D_{fst}.

---

[11] Throughout this App. we use instantaneous growth rates such as $\dot{p}/p$ rather than discrete growth rates $(p_t - p_{t-1})/p_{t-1}$. For the latter the propositions do not appear to hold as stated.

Let us use an asterisk to denote percentage growth rates, so that for a function f, $f^* \equiv \dot{f}/f$. As before, let B_{ft} represent the base-year measure of productivity for the economy as a whole, while B_{fst} is its equivalent for sector s, with D_{ft} and D_{fst} interpreted similarly, and let $w_t = \Sigma\, w_{kt}x_{kt}/\Sigma\, w_{k0}x_{kt}$ and $p_t = \Sigma\, p_{it}y_{it}/\Sigma\, p_{i0}y_{it}$ be the Paasche indices of input and output prices, respectively.

Next, we prove, using (12),

Proposition: $D^*_{fst} = D^*_{ft} = B^*_{ft} = \Sigma_s\, \phi_{st}B^*_{fst} + \Sigma_s\, \dot{\theta}_{st}B_{fst}/B_{ft}$, where $\phi_{st} = \Sigma_{i\in s}\, p_{i0}y_{it}/\Sigma_{\forall j}\, p_{j0}y_{jt}$ is sector s's share of the (base-year) value of the economy's output, and $\theta_{st} = \Sigma_{k\in S}\, w_{k0}x_{kt}/\Sigma_{\forall k}\, w_{k0}x_{kt}$ is s's share of the value of the economy's inputs. In other words, the base-year measure of physical productivity growth for the economy is a weighted average of that growth rate, B^*_{fst}, for the various sectors (each sector weighted by its share of the economy's output valued at base-year prices) plus the sum of the $\dot{\theta}_{st}$, the absolute rate of growth of sector s's share of the economy's inputs, multiplied by B_{fst}/B_{ft}, the relative productivity of that sector.

Proof: $B_{ft} = \Sigma\, p_{i0}y_{it}/\Sigma\, w_{k0}x_{kt}$ or (dividing and multiplying the terms corresponding to the outputs of sector s by $\Sigma_{k\in s}\, w_{k0}x_{kt}$),

$$B_{ft} = \sum_{s}\sum_{i\in s}\left(\frac{\sum\limits_{k\in s} w_{k0}x_{kt}}{\sum\limits_{\forall k} w_{k0}x_{kt}}\right)\frac{p_{i0}y_{it}}{\sum\limits_{k\in s} w_{k0}x_{kt}},$$

or

$$B_{ft} = \sum \theta\; P \qquad (15)$$

Differentiating with respect to t, we

$$\dot{B}_{ft} = \sum_{s} \theta$$

so that

$$B^*_{ft} = \sum_{s} \frac{\theta_{st}B_{fst}}{B_{ft}}\, B_{fst}$$

But

$$\frac{\theta_{st}B_{fst}}{B_{ft}} = \frac{\sum\limits_{k\in S} w_{k0}x_{kt}}{\sum\limits_{\forall k} w_{k0}x_{kt}}\;\frac{\sum\limits_{i\in s} p_{i0}y_{it}}{\sum\limits_{k\in s} w_{k0}x_{kt}}\;\frac{\sum\limits_{\forall k} w_{k0}x_{kt}}{\sum\limits_{\forall j} p_{j0}y_{jt}} = \frac{\sum\limits_{i\in s} p_{i0}y_{it}}{\sum\limits_{\forall j} p_{j0}y_{jt}} = \phi_{st}.$$

so that

$$B^*_{ft} = \sum_{s} \phi_{st}B^*_{fst} + \sum_{s} \frac{\dot{\theta}_{st}B_{fst}}{B_{ft}}. \qquad (16)$$

Q.E.D.

As we have seen in the text, B_{fst}/B_{ft} is not to be interpreted as a measure of the relative productivity *level* of sector s, but as a measure of its relative *growth rate* since the base period. Thus, (16) tells us if there is an increase in the share of output, ϕ_{st}, of the sectors with a currently rapid rate of physical productivity growth, B^*_{fst}, or if there is an increase in the rate of growth in the share of the input value, $\dot{\theta}_{st}$, of the sectors with outstanding recent growth records, that will contribute to B^*_{ft}, which, by (12), is equal to D^*_{ft}. That is how physical productivity growth in the various sectors is translated into the uniform value growth rate, $D^*_{ft} = (w_t/p_t)^*$, for every sector of the economy.

References

Baily, Martin N. "The Productivity Growth Slowdown by Industry." *Brookings Papers Econ. Activity*, no. 2 (1982), pp. 423–54.

Denison, Edward F. "The Shift to Services and the Rate of Productivity Change." *Survey Current Bus.* 53 (October 1973): 20–35.

——— *Accounting for Slower Economic Growth: The United States in the 1970's.* Washington: Brookings Inst., 1979. (*a*)

———. "Explanations of Declining Productivity Growth." *Survey Current Bus.* 59, no. 8, pt. 2 (August 1979): 1–24. (*b*)

———. "Accounting for Slower Economic Growth: An Update." Mimeographed. Washington. Brookings Inst., 1982.

Diewert, W. E. "Aggregation Problems in the Measurement of Capital." In *The Measurement of Capital*, edited by Dan Usher. Chicago: Univ. Chicago Press (for N.B.E.R.), 1980.

Gollop, Frank M. "Evidence for a Sector-biased or Sector-neutral Industrial Strategy: Analysis of the Productivity Slowdown." Mimeographed. Boston: Boston College, Dept. Econ., 1982.

Gollop, Frank M., and Jorgenson, Dale W. "U.S. Productivity Growth by Industry, 1947–73." In *New Developments in Productivity Measurement and Analysis*, edited by John W. Kendrick and Beatrice N. Vaccara. Chicago: Univ. Chicago Press (for N.B.E.R.), 1980.

Kutscher, Ronald E.; Mark, Jerome A.; and Norsworthy, John R. "The Productivity Slowdown and Outlook to 1985." *Monthly Labor Rev.* 100 (May 1977): 3–8.

Nordhaus, William D. "The Recent Productivity Slowdown." *Brookings Papers Econ. Activity*, no. 3 (1972), pp. 493–536.

Norsworthy, John R., Harper, M. J.; and Kunze, Kent. "The Slowdown in Productivity Growth: Analysis of Some Contributing Factors." *Brookings Papers Econ. Activity*, no. 2 (1979), pp. 387–421.

Stigler, George J. *Trends in Output and Employment.* Twenty-fifth Anniversary ser., no. 4. New York: N.B.E.R., 1947.

Thurow, Lester. "The Productivity Problem." Mimeographed. Cambridge, Mass.: MIT, Dept. Econ., November 1980.

Price Leadership and Dynamic Aspects of Oligopoly in U.S. Manufacturing

Micha Gisser

University of New Mexico

This study, which covers a sample of 314 four-digit industries taken from an exhaustive set of 450 four-digit industries, shows that increases in concentration associated with rising productivity occur mainly in low-concentration industries, while decreases in concentration associated with rising productivity occur mainly in high-concentration industries. The empirical results of this study indirectly lend support to a plausible hypothesis that a small group of firms makes a big impact on the productivity of an initially unconcentrated industry and thereby concentrates it. Later, the small firms imitate the now big firms, and concentration goes down while productivity keeps rising.

I. Introduction

No field in the industrial orga
of welfare losses due to oligop
writes that "the reader has by
mand strong justification for
justification for entering is tha
concentration that are associated
mainly in low-concentration industries, while decreases in
tion that are associated with rising productivity occur mainly in high-concentration industries. This empirical finding may be explained by an argument that a small group of innovative and efficient firms makes a big impact on the productivity of an initially unconcentrated

---

I am indebted to an anonymous reader, George J. Stigler, and Thomas H. Goodwin for useful comments. Responsibility for errors is all mine.

[*Journal of Political Economy*, 1984, vol. 92, no. 6]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9206-0004$01.50

industry and thereby concentrates it. Later, the small firms imitate the now big firms, and concentration goes down while productivity keeps rising.

In order to estimate empirically the link between concentration and efficiency, I extend my study of U.S. food manufacturing (Gisser 1982) to include all U.S. manufacturing, covering a sample of 314 four-digit industries taken from an exhaustive set of 450 four-digit industries. Demsetz (1973), Peltzman (1977), Lustgarten (1979), and Kendrick and Grossman (1980) demonstrated that positive changes in concentration ratios in manufacturing industries are associated with increased efficiency. Empirically I confirm their results by utilizing a larger sample of industries. I differentiate my product, however, by utilizing a price leadership model in order to shed light on the different empirical results obtained for low-concentration industries as compared with high-concentration industries.

To anticipate, my basic empirical results are that the increased productivity is greater if the increased concentration comes in a low-concentration industry. Declining concentration enhances productivity too, but only if it occurs in a highly concentrated industry.

II. Measuring Efficiency Gains and the Dynamic Aspects of Oligopoly

A. Concentration and Productivity

Following Solow (1957), I apply a simple procedure for measuring total factor productivity (TFP) involving a neutral shift in technology:

$$\frac{\dot{Q}}{Q} = \frac{\dot{A}}{A} + \alpha\left(\frac{\dot{L}}{L}\right) + (1 - \alpha)\left(\frac{\dot{K}}{K}\right), \tag{1}$$

where Q is output; L and K denote labor and capital, respectively; $\dot{A}/A$ measures the neutral shift in the relevant production function; and α is the share of labor in the production costs. In order to test the hypothesis that productivity is associated with concentration, the following equation is offered for an empirical test:

$$\frac{\dot{A}}{A} = F\left(\frac{\dot{Q}}{Q}, D_1\dot{C}, D_2\dot{C}, C\right), \tag{2}$$

where $\dot{A}/A$ is given by equation (1); $\dot{Q}/Q$ is the rate of output growth; C is the concentration ratio; D_1 is a dummy variable that assumes a value of one for increasing concentration ratios, zero otherwise; and D_2 is a dummy variable that takes on a value of one for decreasing concentration ratios, zero otherwise. Although scale economies are absent in equation (1), following Lustgarten (1979), I include $\dot{Q}/Q$ as an exogenous variable in equation (2). The motive for the inclusion is that the presence of scale economies makes productivity changes a function of real output growth. However, since the inclusion of $\dot{Q}/Q$ raises serious statistical problems,[1] regressions are run once with $\dot{Q}/Q$ and once without. Indeed, the omission of $\dot{Q}/Q$ improves the estimate of the regression coefficient of $D_2\dot{C}$ in table 1.

B. The Data

The nature of the available data forced me to settle for the period 1963–72. Of the total 450 four-digit industries, 135 industries were dropped because of changes in classification between 1963 and 1972, or for similar reasons. U.S. Bureau of the Census (1976, "Concentration Ratios in Manufacturing," table 5) provided the basis for determining which industries to drop.

A detailed discussion of the raw data and how they were converted into input data for the regressions is available in Appendix A.

C. The Statistical Association between Concentration and Productivity

Kendrick and Grossman (1980) regressed productivity growth in 20 manufacturing industries on variables such as the concentration ratio, unionism, and expenditures on R & D. Omitted variables, such as unionism and R & D, are worrisome only [illegible] with included variables. Kendrick [illegible] and R & D are each positively co[illegible] Note, however, that the regressio[illegible] in the Kendrick and Grossman [illegible] unionism probably biases the co[illegible] tion. The regression coefficient o[illegible] man regression is positive; thus my [illegible] the coefficient of C in the positive direction [illegible] of unionism and R & D significantly biases either D_1C or D_2C in my study.[2]

The results for 314 industries contained in table 1 confirm, first, the hypothesis that in manufacturing industries increases in concentra-

---

[1] Spurious correlation between $\dot{Q}/Q$ and $\dot{A}/A$ may exist. There is also the problem of incorrect identification. For a more detailed discussion, see Lustgarten (1979, p. 187).

[2] I calculated the correlation coefficients between C and $D_1\dot{C}$ and C and $D_2\dot{C}$ at .05 and −.16, respectively. Consider, e.g., the impact of omitting R & D on the regression coefficient of $D_1\dot{C}$. Let 1, 2, and 3 stand for $D_1\dot{C}$, C, and R & D. Then the inequality $0 \leq r_{12}^2 + r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23} \leq 1$ places an upper bound on r_{13}, the higher r_{23}, the lower is the upper bound. In the limiting case where r_{23} is unity, r_{13} is equal to r_{12}, which is .05 in this case.

TABLE 1

REGRESSION ESTIMATES OF EQUATION (2),* 1963–72, 314 INDUSTRIES

Regression		Constant	$\dot{Q}/Q$	$D_1\dot{C}$	$D_2\dot{C}$	C	F	R^2
					Labor Productivity			
1	Coefficient	748	337	1.115	074	017	78.182	503
	t-ratio	3.240	15.379	6.152	308	3.464		
2	Coefficient	1.591		1.227	−.561	.023	14.493	.123
	t-ratio	5.345		5.107	−1.777	3.529		
					Total Factor Productivity			
3	Coefficient	.246	308	.854	.011	014	48.564	386
	t-ratio	.943	12.447	4.171	.041	2.465		
4	Coefficient	1.016		956	−.569	0.019	8.742	078
	t-ratio	3.278		3.823	−1.732	2.820		

* $\dot{A}/A = a_0 + a_1(\dot{Q}/Q) + a_2(D_1\dot{C}) + a_3(D_2\dot{C}) + a_4C$

tion ratios are associated with increases in factor productivity and, second, the hypothesis that concentration ratios are associated with increases in factor productivity. In all four regressions, the coefficient of $D_1\dot{C}$ is significant (one tail) at the 0.5 percent level or better, and the coefficient of C is significant at the 1 percent level or better. The impact of omitting $\dot{Q}/Q$ (regressions 2 and 4) upon the coefficients of $D_1\dot{C}$ and C is relatively small; but the coefficient of $D_2\dot{C}$ switches from positive and insignificant to negative and significant at the 5 percent level. When labor productivity is replaced by total factor productivity, the regression coefficient of $D_1\dot{C}$ declines slightly from 1.115 in regression 1 to 0.854 in regression 3. The regression coefficient of C follows a similar pattern: it falls slightly from 0.017 to 0.014. The significance of the regression coefficient of C is somewhat surprising; this coefficient is hardly significant when the same test is applied to food manufacturing data (Gisser 1982). The decline of the regression coefficients of $D_1\dot{C}$ and C when total factor productivity replaces labor productivity may be explained by the fact that the process of discovering new technologies absorbs additional capital, or that it is labor saving, or both.

Peltzman (1977) and Lustgarten (1979) estimated the effects of changes in concentration on changes in unit costs. I confirm their results by estimating the effects of changes in concentration on changes in factor productivity. In so doing I follow Kendrick and Grossman (1980) with two exceptions: my data refer to four-digit industries rather than two, and I have added to their regression the rate of change in concentration as an independent variable.

D. *The Dynamics of Oligopoly: A Plausible Hypothesis and Some Evidence*

I would like to test the following plausible hypothesis: A small number of firms, known as the leaders, invent a relatively *more efficient* process of production. Their marginal cost curve shifts downward faster than the supply curve of the rest of the industry. As a result, with the passage of time, their share in total industry output increases, thus causing the concentration ratio to rise and leading to higher profits for the leaders. This raises the appetite of the other firms, known as the price takers, which by now have had time to decipher the technological secrets of the leaders and adapt them to their own processes of production. Their supply curve now takes a turn in outstripping the decline of the leaders' marginal cost, thus causing the concentration ratio to assume a falling trend.[3]

If this hypothesis is plausible, then we should expect the statistical association between $\dot{A}/A$ and $D_1\dot{C}$ to be positive and relatively strong in industries with low concentration ratios and relatively weak in industries with high concentration ratios. Conversely, the statistical association between $\dot{A}/A$ and $D_2\dot{C}$ would be expected to be negative and relatively weak in industries with low concentration ratios and relatively strong in industries with recent increases in C, approximated by industries with high concentration ratios. To test the hypothesis, I partitioned the sample of 314 industries into two subsamples: first 256 industries with $C \leq 60$ percent, and second 58 industries with $C >$ 60 percent. The results of estimating [illegible]
separately are shown in table 2 [illegible]
$D_1\dot{C}$ for the low-concentration [illegible]
significant with t-ratios of 1.221 [illegible]
respectively. For the group with h[illegible]
t-ratios drop to 1.527 and 0.705 [illegible]
The regression coefficient of D[illegible]
concentration group with t-ratios of 1 [illegible]
1 and 2, respectively. Moreover, in regression [illegible]
has the wrong sign. It has the correct sign in both regressions 3 and 4 and although its t-ratio in regression 3 is only -1.524, it rises to -2.114 in regression 4. While such a statistical test is far from conclu-

---

[3] The plausible hypothesis contains some Schumpeterian elements (1949, p. 133), but I stopped short of calling it a Schumpeterian hypothesis because, as pointed out to me by George Stigler and an anonymous reader, there is no direct proof that small firms copy the innovations of large firms. Data based on four-digit industries allow only an indirect inference; to test rigorously the Schumpeterian hypothesis one should study separately for each industry the cycles and associated set of dates.

TABLE 2

REGRESSION ESTIMATES OF EQUATION (2),* 1963–72; TOTAL FACTOR PRODUCTIVITY FOR HIGH- AND LOW-CONCENTRATION INDUSTRIES

Regression		Constant	$\dot{Q}/Q$	$D_1\dot{C}$	$D_2\dot{C}$	C	F	R^2
			$C \leq 60$ (256 Industries)					
1	Coefficient	.326	.295	.854	319	.016	44.480	.415
	t-ratio	1.113	11.675	4 221	1.123	1.854		
2	Coefficient	.897		1 000	−.130	027	9.025	.097
	t-ratio	2 508		3 997	−.371	2.616		
			$C > 60$ (58 Industries)					
3	Coefficient	−3.779	353	1.237	−1.237	.057	7.449	360
	t-ratio	−1.457	4 560	1 527	−1.524	1.708		
4	Coefficient	−2.143		.660	−1.965	.054	2.197	.109
	t-ratio	−.714		.705	−2.114	1.391		

NOTE.—Chow tests $F(5,304)$, regression 3, table 1 and regressions 1 and 3 table 2 2.013, $F(4,306)$, regression 4, table 1 and regressions 2 and 4, table 2 2.550

* $\dot{A}/A = a_0 + a_1(\dot{Q}/Q) + a_2(D_1\dot{C}) + a_3(D_2\dot{C}) + a_4C$

sive, it nevertheless lends support to what might be called the dynamic hypothesis of oligopolies.[1]

The Chow tests presented at the bottom of table 2 reveal that the null hypothesis of no structural shift between the low-concentration industries and the high-concentration industries can be rejected at the 5 percent level when $\dot{Q}/Q$ is omitted from the regressions and at the 10 percent level when $\dot{Q}/Q$ is included. These results render the previous conclusions more robust.

III. The Price Level and the Dynamic Aspects of Oligopoly

In the short run, a rise in productivity associated with increasing concentration leads to a downward shift in the aggregate marginal cost curve of the price leaders, and it must result in a falling price.[5] In the long run this is true only if the monopoly power is not increased sufficiently. Thus, to confirm the hypothesis that increasing concentration is associated with an increase in productivity, we must show a strong and negative relationship between rising concentration and price changes.[6] Following Eckard (1981), I estimate the contribution of industry i to a change in price level by R_i as

---

[1] The selection of the 60 percent cutoff point is arbitrary. I did not select the more traditional 70 percent cutoff point in order not to exclude some industries whose concentration ratios might have reached a peak in the 60–70 percent range. The results of selecting a 70 percent cutoff point are contained in App. B. When $\dot{Q}/Q$ is included in regressions 1 and 3, the results are as expected, except that the highly concentrated industries come out with significant coefficients for both $D_1\dot{C}$ and $D_2\dot{C}$. When $\dot{Q}/Q$ is omitted in regressions 2 and 4, the results emerge far more robust than in table 2.

[5] Falling prices are not the only consequence of increased productivity associated with increased concentration. Another possible consequence could be a shift in demand due to improved product. Unlike falling prices, this would be difficult, if not impossible, to quantify.

$$R_i = \frac{VA_i^1/q_i^1}{VA_i^0/q_i^0} = \frac{VA_i^1/VA_i^0}{q_i^1/q_i^0}. \tag{3}$$

VA_i^1/VA_i^0 is the ratio of the value added for good i at the end of year 1 to the base year. The data for value added were obtained from the U.S. Bureau of the Census (1966; 1976, table 3). In order to free the measure R_i from the general inflation, this ratio is divided by the increase in the implicit price deflator between 1963 and 1972. The ratio q_i^1/q_i^0 is calculated from production indices the derivation of which is set forth in Appendix A. In order to test the hypothesis that increasing concentration is associated with falling prices, I denote the annual percentage change in R by $\dot{R}/R$ and offer the following equation for an empirical test:

$$\frac{\dot{R}}{R} = F\left(\frac{\dot{Q}}{Q}, D_1\dot{C}, D_2\dot{C}, C\right). \tag{4}$$

The results are presented in table 3. First note that the coefficients of both $D_1\dot{C}$ and $D_2\dot{C}$ are significant in both [illegible] coefficient of $D_1\dot{C}$ is negative and consistently value. This suggests that an incre age point is expected to result in a productivity increases translate in The coefficient of $D_2\dot{C}$ is positive, from the regression. Since I am

---

[6] Such a negative relationship between rising concentration and the price level will also shed light on the lively debate concerning the impact of concentration on inflation Eckard (1981) investigated empirically the relationship between concentration changes and inflation. His study covers the periods 1958 63, 1963–67, and 1967–72 My study covers the period 1963–72 By employing various statistical procedures other than regression analysis, Eckard shows that a positive relationship exists between negative concentration changes and inflation and a negative relationship exists between positive concentration changes and inflation. By applying regression analysis to my sample of 314 four-digit industries, I am able not only to lend support to Eckard's conclusions, but also to show that industries with a relatively high level of concentration (C) contribute to a significant reduction in the price level It is worth noting that Eckard wrote his article in response to a theoretical article in which Johnson and I showed that, ceteris paribus, an increase in concentration would result in only a minute rise in the price level (Gisser and Johnson 1979). Our ceteris paribus covered the state of the art (technology), among other things. We have not been in any disagreement with Eckard.

TABLE 3

REGRESSION ESTIMATES OF EQUATION (4),* 1963–72, 314 INDUSTRIES

Regression		Constant	$\dot{Q}/Q$	$D_1\dot{C}$	$D_2\dot{C}$	C	F	R^2
1	Coefficient	1.040	−.207	−.586	.534	−.023	39.266	.337
	t-ratio	4.558	−9.582	−3.271	2.240	−4.906		
2	Coefficient	.521		−.654	.925	−.027	16.822	.140
	t-ratio	2.069		−3.217	3.462	−4.989		

* $\dot{R}/R = a_0 + a_1(\dot{Q}/Q) + a_2(D_1\dot{C}) + a_3(D_2\dot{C}) + a_4C$

correctly specified, I can only conclude that this coefficient is positive and significant and that its value lies somewhere between 0.5 and 1.0.[7]

The results of testing the plausible hypothesis are contained in table 4. The regression coefficients of $D_1\dot{C}$ for the low-concentration group are negative as expected and highly significant, with *t*-ratios of −2.874 and −2.967 in regressions 1 and 2, respectively. For the group with high concentration the values of the *t*-ratios drop to −1.776 and −1.489, respectively. The regression of $D_2\dot{C}$ is insignificant for the low-concentration group with *t*-ratios of 0.438 and 1.530 in regressions 1 and 2, respectively. It has the expected positive sign in regressions 3 and 4, a robust result as indicated by *t*-ratios of 3.644 and 3.994, respectively.

The Chow tests presented at the bottom of table 4 reveal that the null hypothesis of no structural shift between the low-concentration industries and the high-concentration industries can be rejected at the 1 percent level. Thus the regression estimates in this section lend additional support to the plausible dynamic hypothesis, which was confirmed by testing a different model in the previous section.

IV. Summary and Conclusions

A sample of 314 four-digit manufacturing industries was used in order to investigate the statistical association between concentration and efficiency. By estimating Solow-type technological changes be-

---

[7] A surprising result is that the coefficient of C is roughly equal to −0.025 and is statistically more significant than the coefficients of $D_1\dot{C}$ and $D_2\dot{C}$. It implies that industries that have reached a high level of concentration devote more effort to discovery of new cost-reducing technologies than industries with low concentration levels. Eckard concludes: "In fact, the evidence suggests that neither increasing nor relatively high concentration is associated with 'bad' inflationary performance. Policymakers will have to look elsewhere for a cure for inflation" (1981, p. 1050). My conclusion is stronger, i.e., that policymakers may derive comfort from the unambiguous regression results that suggest that both increasing and relatively high concentration are associated with "good" inflationary performance.

TABLE 4

REGRESSION ESTIMATES OF EQUATION (4),* 1963–72 FOR HIGH- AND LOW-CONCENTRATION INDUSTRIES

Regression		Constant	$\dot{Q}/Q$	$D_1\dot{C}$	$D_2\dot{C}$	C	F	R^2
				$C \leq 60$ (256 Industries)				
1	*Coefficient*	1.121	−.222	−.511	.109	−.030	37.450	.374
	t-ratio	4.354	−9.981	−2.874	.438	−4.054		
2	Coefficient	.691		−.621	.447	−.039	12.020	.125
	t-ratio	2.309		−2.967	1.530	−4.458		
				$C > 60$ (58 Industries)				
3	Coefficient	5.391	−.110	−1.156	2.376	−.074	6.884	.341
	t-ratio	2.586	−1.769	−1.776	3.644	−2.764		
4	Coefficient	4.880		−.976	2.603	.074	7.774	.304
	t-ratio	2.319		−1.489	3.994	−2.680		

NOTE.—Chow tests *F*(5,304), regression 1, table 3 and regressions 1 and 3, table 4: 4.332; *F*(4,306) regression 2, table 3 and regressions 2 and 4, table 4: 1.306.

* $\dot{R}/R = a_0 + a_1(\dot{Q}/Q) + a_2(D_1\dot{C}) + a_3(D_2\dot{C}) + a_4C$.

tween 1963 and 1972, I was able to confirm what was shown earlier by Peltzman (1977), Lustgarten (1979), and Kendrick and Grossman (1980), namely, that increasing concentration is associated with increased total input productivity. I obtained better results for labor productivity than for total factor productivity, an indication that efficiency does not come free, and to
saving.

By investigating low-concentra
concentration industries, I was
dynamic hypothesis that initially
opment of new technologies and
manage to increase their market
appetite of the rest of the price takers
to learn, they emulate the leaders and
market. The implication of this dynamic model is that supernormal profits constitute an incentive for technological innovation rather than a mere welfare loss to society.

The dynamic hypothesis above is summarized by figure 1. D_0 is the market demand, and S is the supply of the small firms. D'' is the net demand curve confronting the innovating leaders in the industry. In absence of any association between concentration and efficiency, and if for simplicity collusion among the leaders is assumed, the price would be set at P_m and the concentration ratio would be $\overline{BC}/\overline{AC}$ in figure 1*a*. But if concentration and efficiency go hand in hand, the intertemporal shift of the aggregate marginal cost curve of the lead-

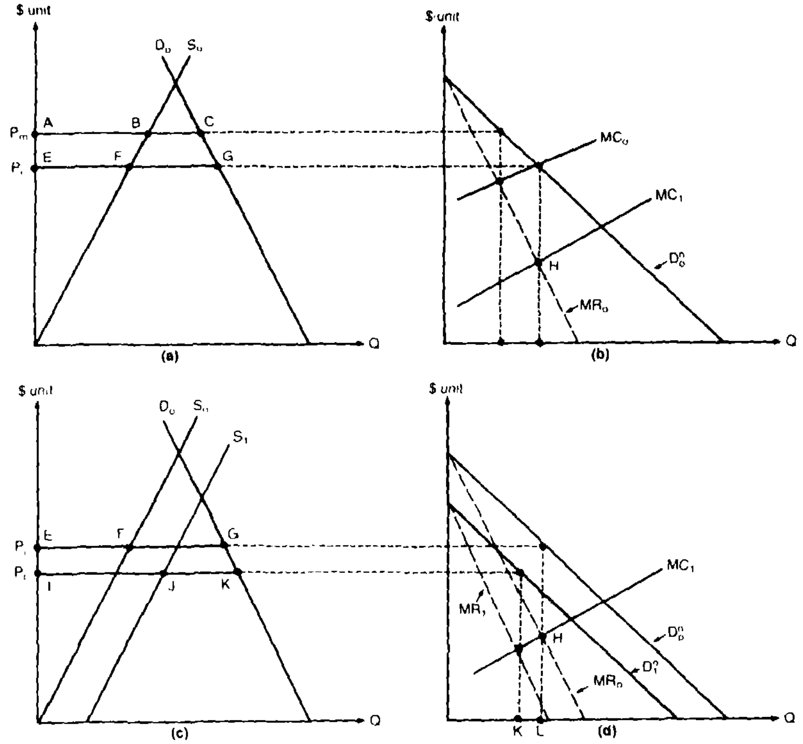


FIG. 1.—A geometric illustration of the plausible hypothesis. Curves are drawn as linear for geometric convenience and clarity. All curves are long run

ers who invest in a new technology exceeds the intertemporal shift of the supply curve of the small firms. This is depicted in figure 1*b* by a net shift from MC_0 to MC_1. Consequently, the concentration ratio rises to $\overline{FG}/\overline{EG}$ ($\overline{EF} < \overline{AB}$). It is not unlikely that, as shown in 1*b*, the shift in the marginal cost of the leaders would be sufficient to restore a competitive price P_c, or even a lower price.[8] The second phase in the cycle occurs when the small firms copy the technology of the leaders: they then benefit from a net intertemporal shift of their supply curve from S_0 to S_1 in 1*c*. As a result, the leaders' net demand decreases from D_0'' to D_1'', the price further falls to P_1', and the concentration ratio reduces to a level $\overline{JK}/\overline{IK}$ (the share of the leaders declines by $\overline{KL}$ in 1*d*).

---

[8] Consider an average industry with an aggregate demand and marginal cost curves of unitary elasticity Assuming a Cobb-Douglas production function the aggregate marginal cost function of the leaders is described by $Q = GK^{(1+\epsilon)}(MC)^{\epsilon}$, where K is the production-function shifter. If the gains from concentration efficiency are to restore the competitive price P_c, K must increase until the marginal cost curve of the leaders shifts downward to position MC_1 in panel *b* of fig. 1, where it cuts the leaders' marginal revenue curve at point H. Utilizing the above equation of the MC curve, it can be shown that to shift the marginal cost curve from MC_0 to MC_1 K must increase from 1 to 1.1486. During the 9 years under study, the concentration ratio increased at an annual rate of 0.73 for the 157 industries with a positive $\dot{C}$. This means that from 1963 to 1972, for

Employing the same sample in order to study the impact of concentration on the price level provided an alternative empirical confirmation of the hypothesis that concentration and efficiency are positively correlated. Moreover, it confirmed the dynamic hypothesis of oligopolistic behavior with more robust results.

My major conclusion is that antitrust agencies should not attempt to restrict concentration by restructuring manufacturing industries. Such attempts might deprive society of the apparent benefits associated with intertemporal changes in the concentration of these industries.

Appendix A

The Data

Annual percentage changes in output ($\dot{Q}/Q$) were calculated from cross-weighted production indices taken from U.S. Bureau of the Census (1973, table 3, 1977, table 1*c*). These data, I admit, are not of the highest quality mainly because the census people estimate output by deflating the value of shipment by various price indices that imperfectly match the SIC four-digit industry definitions. This point was discussed [illegible] (1970). The scope of my study necessitates the use
data do not exist at the four-digit le

Annual percentage changes in la
data obtained from U.S. Bureau of
table 3; 1976, "General Summary," t
of man-hours provided by produ
changes of nonproduction employe
average of the two was calculated.
the ratio of the total wage bill of pro
base year and end of year and then a

---

these industries the concentration ratio rose 6.57 percentage points. In table 1 the smallest coefficient of $D_1\dot{C}$ is 0.854 in regression 3. Although this coefficient may be viewed as a weighted average of price makers and price takers, the positive correlation between $D_1\dot{C}$ and $\dot{A}/A$ lends strong support to the hypothesis that most of the productivity improvement originates with the price leaders. Given their 39 percent average share in total output, it can be shown that each percentage-point increase in concentration is associated with a 2.20 percent increase in the productivity of the leaders. It follows that an increase of 6.57 percentage points in the concentration ratio is associated with a 14.45 percent shift in the leaders' production function, roughly what is needed to restore the competitive price.

TABLE A1

MEANS AND STANDARD DEVIATIONS, 314 OBSERVATIONS

Variable	Mean	Standard Deviation
$\dot{Q}/Q$	3.831	4.630
$\dot{L}/L$	.746	3.601
$\dot{K}/K$	4.258	5.426
$\dot{A}/A$ (labor)	3.085	2.475
$\dot{A}/A$ (total)	2.264	2.513
$D_1\dot{C}$	.364	.595
$D_2\dot{C}$	−.278	.458
C	39.250	20.880
$\dot{R}/R$	−1.041	2.116

Annual percentage changes in capital input ($\dot{K}/K$) were derived from data taken from the *Annual Survey of Manufactures* (1968, chaps. 6, 7, and 1978). The value of capital input was estimated by adding 10 percent of gross book value of depreciable assets to expenditures for new plant and new equipment.[9] Figures for 1972 were deflated by the implicit GNP deflator for gross private fixed nonresidential investment.

Annual percentage shifts in the production function ($\dot{A}/A$) were calculated by subtracting the weighted average of $\dot{L}/L$ and $\dot{K}/K$ from $\dot{Q}/Q$, as shown in equation (1). Figures provided by Kendrick and Grossman (1980, tables 2–3, p. 27) were utilized to estimate the weights, denoted by α. These figures are available only at the two-digit SIC.

Concentration ratio (C) is defined as the share of shipment value accounted for by the four largest companies. Data were obtained from U.S. Bureau of the Census (1976, "Concentration Ratios in Manufacturing," table 5), and given as percentages. C is the average of the concentration ratios in the base year and the end-of-period year.

Change in concentration ratio ($\dot{C}$) is defined as the change in concentration over the change in time (in years). It can be expressed as $\dot{C} = (\Delta C)/(\Delta t)$. As an illustration, for SIC 2011 the concentration ratio fell from 31 percent in 1963 to 22 percent in 1972. Hence $\dot{C}$ was calculated as $(22 - 31)/9 = -1$. Census concentration data suffer mainly by ignoring imports and by assuming national market definitions. Schwartzman and Bodoff (1971) compared the weighted average local concentration ratios of local industries with corresponding national concentration ratios in 1963. They found that the average local concentration ratio in regional manufacturing industries exceeds the national concentration ratio. When they adjusted the national average of 37 percent by substituting the regional and local concentration ratios for the national concentration ratios of regional and local industries it increased by 5 percentage points. This means that the data for concentration (C) are somewhat flawed, but the bias cancels out in the data for the change in concentration ($\dot{C}$). This is not correct if imports have been changing rapidly over the period (e.g., automobiles) or if technology has changed the extent of the geographic market over the period (e.g , beer). The failure to consider exports causes similar problems.

---

[9] Information about rental payments made during the year is a legitimate component of capital input. I opted not to utilize it because it was not available in about half of the four-digit industries. In 37 industries, capital input does not include 10 percent of gross book value of depreciable assets because the data were lacking.

Appendix B

REGRESSION ESTIMATES OF EQUATION (2),* 1963–72, 314 INDUSTRIES

Regression		Constant	$\dot{Q}/Q$	D_1C	$D_2\dot{C}$	C	F	R^2
				$C \leq 70$ (286 Industries)				
1	Coefficient	.380	307	755	294	013	42.920	379
	t-ratio	1 336	11 762	3 619	1 019	1.783		
2	Coefficient	1.050		916	210	021	7 476	074
	t-ratio	3.089		3 609	602	2.464		
				$C > 70$ (28 Industries)				
3	Coefficient	−4 121	354	3.132	−2 468	047	9 347	619
	t-ratio	− 980	1.452	3 050	−3 090	920		
4	Coefficient	−1 543		1 232	−3 077	038	3 283	291
	t-ratio	277		987	2 928	567		

* $\dot{A}/A = a_0 + a_1(\dot{Q}/Q) + a_2(D_1C) + a_3(D_2\dot{C}) + a_4C$

References

Demsetz, Harold. "Industry Struct
J. Law and Econ. 16 (April 1973)
Eckard, E. Woodrow, Jr. "Concent
dence." *J P E* 89 (October 1981)
Gisser, Micha. "Welfare Implicatio
ing." *American J Agricultural Ec*
Gisser, Micha, and Johnson, Ronald
tion." *J.P E.* 87 (December 1979)
Kendrick, John W , and Grossman, Elliot
Trends and Cycles Baltimore: Johns Hopkins Univ Press 1980
Lustgarten, Steven "Gains and Losses from Concentration. A Comment" *J. Law and Econ.* 22 (April 1979): 183–90
Peltzman, Sam. "The Gains and Losses from Industrial Concentration" *J. Law and Econ.* 20 (October 1977): 229–63
Scherer, F. M. "The Causes and Consequences of Rising Industrial Concentration." *J Law and Econ.* 22 (April 1979): 191–208.
Schumpeter, Joseph A. *The Theory of Economic Development An Inquiry into Profits, Capital, Credit, Interest, and the Business Cycle.* Translated from the German by Redvers Opie Cambridge, Mass Harvard Univ. Press, 1949
Schwartzman, David, and Bodoff, Joan "Concentration in Regional and Local Industries." *Southern Econ. J* 37 (January 1971) 343–48.
Solow, Robert M. "Technical Change and the Aggregate Production Function." *Rev. Econ. and Stats.* 39 (August 1957). 312–20.

U.S. Bureau of the Census. *Census of Manufactures, 1963.* Vol. 1. *Summary and Subject Statistics.* Washington: Government Printing Office, 1966.

———. *Annual Survey of Manufactures, 1964–65.* Washington: Government Printing Office, 1968.

———. *Census of Manufactures and Mineral Industries, 1967.* Special Report Ser. *Indexes of Production.* Washington: Government Printing Office, 1973.

———. *Census of Manufactures, 1972.* Vol. 1. *Subject and Special Statistics.* Washington: Government Printing Office, 1976.

———. *Census of Manufactures and Mineral Industries, 1972.* Vol. 4. *Indexes of Production.* Washington: Government Printing Office, 1977.

———. *Annual Survey of Manufactures, 1976.* Washington: Government Printing Office, 1978.

The New Divisia Monetary Aggregates

William A. Barnett
University of Texas at Austin

Edward K. Offenbacher
Bank of Israel

Paul A. Spindt
Federal Reserve Board

Barnett's Divisia monetary aggregates were derived to be elements
of Diewert's class of superlative quantity index numbers. Relative to
aggregation theory, Barnett's resulting monetary aggregates are
strictly preferable to the official sum
component monetary assets are
pirical tests based on the relev
likewise uniformly favored th
current article compares the D
gates relative to numerous conv
Divisia monetary aggregates, es
usually perform best in these t

We have benefited from the comments of Robert Barro. We also have benefited from the comments on this research of participants at conferences at the N.B.E.R. (Cambridge, 1981), at the European Econometric Society meetings (Athens, Greece, 1979), at the North American Econometric Society meetings (Denver, 1980), at the Latin American Econometric Society meetings (Santiago, Chile, 1983), at the University of Aix-Marseille (Aix-en-Provence, France, 1980), and at the University of Chicago (1980, 1982). Some of the research on which this paper is based was conducted while we were at the Federal Reserve Board. The views expressed herein are solely ours and do not necessarily represent the views of the Board of Governors of the Federal Reserve System. The research on this paper was partially supported by National Science Foundation grant SOC 8305162 along with the Sam P. Woodson Memorial Centennial chair and the Janey Briscoe Centennial fellowship at the University of Texas at Austin.

[*Journal of Political Economy*, 1984, vol. 92, no. 6]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9206-0001$01.50

I. Introduction

Recently a rapidly growing line of research has appeared concerning the rigorous use of aggregation and index number theory in the construction of monetary aggregates based on Diewert's (1976) "superlative" class of quantity index numbers. Research on this application of the superlative class was motivated by and built on Barnett's (1980*a*) proposal and initial results.[1] Barnett (1980*a*, p. 38; 1981*a*, p. 221) proposed use of either the Divisia or Fisher ideal index for monetary quantity aggregation. Since all such index numbers in the superlative class move very closely together (and are in fact usually identical, to within round-off error, for monetary asset data), the choice among elements of the class is little more than arbitrary. The purpose of this article is to provide a quantitative assessment of the relative merits of the official (summation) versus Divisia monetary quantity indexes, constructed with the same components and component groupings. The sample for the comparisons consists of quarterly data from 1959 through the end of 1982.[2]

II. History and Objectives

Although many types of financial intermediaries and monetary substitutes have evolved over the past 30 years, most economists have placed little faith in broad monetary aggregates, since summation aggregation has long seemed inappropriate at high levels of aggregation over imperfect substitutes. Friedman and Schwartz (1970, pp. 151–52) clearly described the problem with the high-level official aggregates:

---

[1] Some of the foundations of this research are contained in Barnett (1978, 1980*a*, 1981*a*, 1982*b*, 1983*a*) and Barnett and Spindt (1979), and an overview of that literature is contained in Barnett, Offenbacher, and Spindt (1981). The first official source of the historical data, supplied for 1969–81, is Barnett and Spindt (1982), which is available on request at no cost from the Publications Services Department at the Federal Reserve Board in Washington. Data for 1979–83 are supplied in the app. of Barnett (1981*b*). The complete data series, continually updated to the latest month, are available to subscribers to the services of Data Resources Incorporated (DRI). Those series are located in DRI data bank USEMS/DATA under the series names JM1D, JM2D, and JM3D for Divisia M1, M2, and M3, respectively. Related research includes Hawtrey (1930), Chetty (1969, 1972); Friedman and Schwartz (1970), Bisignano (1974), Moroney and Wilbratte (1976), Barth, Kraft, and Kraft (1977), Donovan (1978), Offenbacher (1979, 1980); and Barnett (1980*b*). For a unified overview of all that literature, see Barnett (1984*a*). For a discussion of the recent behavior of the Divisia monetary aggregates, see Dorfman (1983).

[2] The original results in Barnett, Spindt, and Offenbacher (1981) used quarterly data ending in 1980. In the present paper we have updated all of the results except those in table 2 and Sec. VI*B*2. The computer programs needed to update those results were unavailable to us when we updated our other results.

> This [summation] procedure is a very special case of the more general approach. In brief, the general approach consists of regarding each asset as a joint product having different degrees of "moneyness," and defining the quantity of money as the weighted sum of the aggregate value of all assets, the weights for individual assets varying from zero to unity with a weight of unity assigned to that asset or assets regarded as having the largest quantity of "moneyness" per dollar of aggregate value. The procedure we have followed implies that all weights are either zero or unity. The more general approach has been suggested frequently but experimented with only occasionally. We conjecture that this approach deserves and will get much more attention than it has so far received.

By equally weighting components, aggregation by summation can badly distort an aggregate. For example, if one wished to obtain an aggregate of transportation vehicles, one would never aggregate by summation over the physical units of, say, subway trains and roller skates. Instead one would construct a quantity index (such as the Department of Commerce's many Laspeyres quantity indexes) using weights based on the *values* of the different modes of transportation. As another example, suppose the money supply were measured by the Federal Reserve's current highest level official aggregate, *L*, which contains most of the national debt of short and intermediate maturity. All of that portion of the national debt could be monetized without increasing either taxes or the "in [illegible] would simply have exchanged [illegible] new Divisia index over the [illegible] transfer as an exchange of "pu[illegible] Divisia *L* would rise at about [illegible] prices that we would expect w[illegible]

The traditionally constructed [illegible] M3) implicitly view distant substitutes [illegible] for currency. Rather than capture only part of the [illegible] tary services, as M1 does, the broad aggregates swamp the included monetary services with heavily weighted nonmonetary services. The result no longer resembles economists' concept of "money." Nevertheless, the need remains for aggregates that capture the contribu-

---

[3] On the other hand, the problems associated with policies that target very low-level aggregates result from the inability of such aggregates to internalize pure substitution effects occurring within the economy's transactions technology, since low level aggregates aggregate over a small subset of the factors of production in that transactions technology.

tions of all monetary assets to the economy's flow of monetary services.

Regarding the simple sum (arithmetic average) index, Irving Fisher wrote over a half century ago that "the simple arithmetic average produces one of the very worst of index numbers, and if this book has no other effect than to lead to the total abandonment of the simple arithmetic type of index number, it will have served a useful purpose. . . . The simple arithmetic [index] should not be used under any circumstances" (1922, pp. 29, 36).

III. The Divisia Monetary Quantity Index

Let m_{jt} and π_{jt} be the quantity and price, respectively, of the jth component of an aggregate during period t. Törnqvist (1936), and subsequently Theil (1967), advocated the following discrete-time approximation to the continuous-time Divisia quantity index:

$$Q_t^* = Q_{t-1}^* \prod_j \left(\frac{m_{jt}}{m_{j,t-1}}\right)^{\frac{1}{2}(s_{jt}+s_{j,t-1})},$$

where $s_{jt} = \pi_{jt}m_{jt}/\sum_k \pi_{kt}m_{kt}$ is the expenditure share on component j during period j.[1] We shall refer to Q_t^* as the Divisia index (in discrete time), although it also frequently is called the Törnqvist index. Taking logarithms of each side, observe that

$$\log Q_t^* - \log Q_{t-1}^* = \sum_j \bar{s}_{jt}(\log m_{jt} - \log m_{j,t-1}), \qquad (1)$$

where $\bar{s}_{jt} = \frac{1}{2}(s_{jt} + s_{j,t-1})$. The same index number results regardless of whether the exact aggregation-theoretic aggregate being approximated is the output of a utility function or of a production function. The aggregation-theoretic procedure for selecting the component assets is described in Barnett (1982*b*).[5]

Diewert (1976) has proved that the Divisia index lies within his class of "superlative" index numbers, which all are nearly identical numerically. In fact, as a quantity index, the Divisia index is by far the most widely used element of Diewert's superlative class, because of the index's numerous theoretical optimality properties. Those remarkable properties result from that index's simultaneous theoretical links with the Divisia line integral, the translog aggregator function, the Malmquist index, and the Konyus index (see, e.g., Diewert 1980, 1981; Caves, Christensen, and Diewert 1982*b*). Recently Caves, Christensen, and Diewert (1982*a*, p. 1411) have proved that the Divisia index is "superlative in a considerably more general sense than shown by Diewert. We are not aware of other indexes that can be shown to be superlative in this more general sense." Also observe that the growth rate of the index is a weighted average of the growth rates of the components. The weights are the share contributions of each component to the total expenditure on the services of all components. Because of the availability of such a natural interpretation and because of the index's optimality properties, we advocate use of the Divisia index to measure the quantity of money at all levels of aggregation, as first proposed by Barnett (1980*a*; 1981*a*, chap. 7).

---

[1] For details regarding the Divisia index, see Barnett (1982*a*). Regarding aggregation over consumers, see Barnett (1981*a*, chap. 3).

[5] The procedure requires testing for blockwise weakly separable groupings of assets. Those tests require knowledge of both the quantity and the user-cost price of each asset that is to be considered as a possible element of a weakly separable group. The formula for the user-cost price of an asset is presented in the next section. A nonparametric approach to testing for blockwise weak separability is available from Pudney (1981).

In order to be able to use (1) for aggregation over monetary asset quantities, we need the price, π_{jt}, corresponding to each component quantity asset, m_{jt}. In economic quantity aggregation, the appropriate price of a component durable good is its user cost. Barnett (1978, 1980*a*) derived the user cost of a monetary asset and found that the current-period user cost, π_{jt}, of m_{it} is

$$\pi_{it} = \frac{p_t^*(R_t - r_{it})(1 - \tau_t)}{1 + R_t(1 - \tau_t)}, \tag{2}$$

where p_t^* is the true cost-of-living index, r_{it} is the own current-period holding yield on component i, R_t is the maximum available expected holding-period yield in the econom [illegible]
The corresponding real user co [illegible]

$$\pi_{it}^{\dagger}$$

The holding period used in de [illegible]
r_{it}, which is a short rate. Details of [illegible]
defined in Barnett and Spindt (1982), [illegible]
used for the r_{it}'s. A few of the more noteworthy details follow. The return on demand deposits is a version of Klein's (1974) competitive rate on demand deposits. The raw data on the other r_{it}'s are for various holding periods. Hence the unadjusted yield differentials, $R - r_{it}$, can reflect differences in term to maturity as well as differences in monetary services at the margin. However, Barnett's derivation of equation (2) requires all yields to be for the same holding period. As a result, all yields are converted to a 1-month holding-period basis by using the Treasury securities yield curve and the yield curve adjustment procedure developed for the Federal Reserve's FRB-Penn-MIT quarterly model (developed jointly by the Federal Reserve Board,

University of Pennsylvania, and MIT). The certainty-equivalence theory on which (2) was based required that any risk premiums be left within the rate structure, if freedom from default risk ("store of value") is to be valued by the Divisia quantity index as a monetary service. We left any such risk premiums within the r_{it}'s, although the components of the existing monetary aggregates all possess very low default risk. If the components were to be selected properly by tests for blockwise weak separability (as described in n. 5), then it is conceivable that assets such as gold or equity shares could appear in the nested aggregates at some level of aggregation. In such cases, the expected holding-period yields, r_{it}, would depend nontrivially on expected capital gains and losses and expected transactions costs.

Observe that the "weights" are not the user costs, but rather the average shares, $\bar{s}_{it}$, which depend jointly on all quantities and user costs. Those share weights were not acquired by an ad hoc weighting scheme (such as weighting by variances, bid-ask spreads, denominations, maturities, velocities, or turnover rates), but rather were derived directly from microeconomic aggregation and index number theory.[6] Also note that in equation (1) the user costs appear only in the share weights, $\bar{s}_{it}$, and all factors in (2) except for $R_t - r_{it}$ cancel out of the numerator and denominator of each $\bar{s}_{it}$. Hence in computing the Divisia monetary aggregates, we could view the user-cost price, π_{it}, as the opportunity cost, $R_t - r_{it}$, which measures the interest forgone by holding monetary asset i when R_t is available. However, an exception to that statement results from the fact that our implicit competitive rate of return on demand deposits is not taxed, while the other rates are. Hence it is not strictly true that all marginal tax rates cancel out of each $\bar{s}_{it}$. We expect that this problem is not of great empirical consequence, but in future work we plan to refine the Divisia monetary aggregates further by using results of recent work on average marginal tax rates, such as Barro and Sahasakul (1983).

IV. Granger Causality and Prediction Risk Reduction

In this section, we report the results of applying several standard tests of the Granger causality relation between money and income, where money is measured first by an official summation aggregate and then by the corresponding Divisia aggregate. We use three different empirical test procedures.

---

[6] See Barnett (1981*b*, 1982*a*). For a general nontechnical discussion of the properties and interpretation of the Divisia monetary aggregates, see Barnett (1983*b*). See Barnett (1984*a*) for a discussion of the various ad hoc weighting schemes (such as the linear regression index, the latent variables index, the so-called Fisher money stock index, M_Q, etc.)

The first procedure is advocated by Pierce and Haugh (1977). In the first two columns of table 1, we display the tail areas for Haugh's (1972) small-sample test against the hypothesis of independence. The results are provided for both 8- and 12-quarter symmetric lags.[7] In every instance, the hypothesis of independence between GNP and money, as measured by either the official or the Divisia aggregate, would be rejected by a test at the 5 percent significance level. However, each Divisia aggregate produced a smaller test tail area than the corresponding official aggregate. As a result, the tail area comparisons favor the Divisia aggregates.

The second approach is due to Sims (1972). The tail areas of the Sims test are presented in the third and fourth columns of table 1 for both directions of causality. We find that in all cases the hypothesis that money does not Granger-cause GNP would be rejected at the 5 percent significance level. However, the tail area for the official (sum) aggregate was less than that for the corresponding Divisia aggregate at each level of aggregation except at the highest level, L. The hypothesis that GNP does not Granger-cause money would be accepted at the 5 percent level with each of the monetary aggregates except for the official M1 aggregate. The tail area for the sum aggregate exceeded that for the corresponding Divisia aggregate at the two highest levels of aggregation, M3 and L.[8]

In the third procedure, we compute an approximate likelihood ratio test of the hypothesis that $C(L) = 0$ in the bivariate autoregression

$$\begin{pmatrix} Z_t \\ Y_t \end{pmatrix} = \begin{bmatrix} A(L) \\ C(L) \end{bmatrix}$$

It is immediate from Granger's
causality) that Z_t does not cause
orders of A, B, C, and D are each
the test of the hypothesis that $C(L)$

---

[7] For details of the estimated ARIMA models, see Barnett, Spindt and Offenbacher (1981)

[8] It is crucial to the Sims procedure to account for autocorrelation of the untransformed disturbances properly We correct for serial correlation in those disturbances by using a general fifth-order autoregressive transformation As a test of the assumed lack of autocorrelation of the resulting transformed disturbances, we computed the tail area of the test that the first eight autocorrelations of the transformed disturbance terms are zero As seen from the tail areas displayed in table A1 in App A, the hypothesis was accepted at the .05 level in all cases

[9] This result follows from Feige and Pearce (1979) Simplifying along the same lines as Feige and Pearce (1979) to avoid the problems discussed by Pierce and Haugh (1977), we rule out instantaneous causality by assuming $b_0 = c_0 = 0$ The model is normalized by taking $a_0 = d_0 = 0$. Using common time-series notation (see, e g., Granger and Newbold 1977, p. 6), A, B, C, and D are polynomials and L is the backshift (lag) operator

TABLE 1

TAIL AREAS OF TESTS OF GRANGER CAUSALITY FROM MONEY MEASURES TO GNP

MONETARY MEASURES	TAIL AREAS OF TEST				
	Haugh-Pierce Test		Sims Test*		
	8-Quarters Lag	12-Quarters Lag	GNP to Money	Money to GNP	Direct Test†
M1	.0000	.0000	.0130	.0038	.074
$M1^D$	.0000	.0000	.1277	.0171	.056
M2	.0222	0113	.1143	0014	.005
$M2^D$	.0105	.0082	1959	0040	.001
M3	0303	.0251	.6950	.0058	.090
$M3^D$	.0267	.0243	.4023	.0088	.021
L	.0152	.0175	0901	0047	.079
L^D	.0016	.0080	.0870	.0029	.002

NOTE.—Sample period, quarterly observations 1959 1–1982 4 The superscript *D* designates a Divisia aggregate The other aggregates are sums.

* Degrees of *F*-statistic = 8, 90

† *Df* of χ^2 statistic = 8

column of table 1.[10] The hypothesis that money does not Granger-cause GNP would be rejected at the 5 percent level with the Divisia aggregates at all levels of aggregation except at the lowest, M1, level, at which the tail area was a marginal .056. The same hypothesis would be accepted at the .05 level with the official summation aggregates at all levels of aggregation except M2. In addition, the tail area of the test for the Divisia aggregates was lower than that for the corresponding sum aggregate at all levels of aggregation. These tail area comparisons favor the Divisia aggregates.

Table 1 does not reveal a single uniformly best aggregate. In terms of 5 percent significance levels, sum M2, Divisia M2, Divisia M3, and Divisia *L* produce successful test results in all cases. In terms of the test tail areas, none of those three aggregates uniformly dominates, although Divisia *L* is most frequently best. While Divisia M1 was unsuccessful in one causality test at the .05 level, the failure was only marginal with a tail area of .056. Otherwise Divisia M1 also did well in these tests.

Because Granger-causality tests are tests of "incremental information content" (Schwert 1979, p. 82), causality results have an important bearing on the usefulness of monetary aggregates as indicators. When a monetary aggregate Granger-causes some variable of policy interest, such as income, the prediction variance of that variable can be reduced by conditioning on measurements of the monetary aggregate. Barnett and Spindt (1979) and Barnett, Spindt, and Offenbacher (1981) have found that the Divis [illegible] results are usually better than the correspond [illegible] cators of a variety of policy targ [illegible] proportionate reduction in pred [illegible] and sum M2 as indicators in the [illegible] terly model are presented in ta [illegible] the generalized variance of the [illegible] variables. In the indicator use of the [illegible] ditional model forecast is revised to a co [illegible] indicator measurement along the lines detailed in Tinsley, Spindt, and Friar (1980, p. 67). It can be seen from table 2 that considerable reductions in prediction risk are achievable by policymakers using Divisia M2 as an indicator.

V. Velocity

A substantial controversy has arisen in the literature regarding a "shift" in the demand-for-money function that frequently is pur-

---

[10] See Wall (1974) for a discussion of the test procedure

TABLE 2

PROPORTIONATE REDUCTION OF PREDICTION RISK WITH RESPECT TO INCOME, PRICES, AND UNEMPLOYMENT: FRB-PENN-MIT QUARTERLY MODEL FORECASTING

MONETARY AGGREGATE	$I_{Y_t\|Z_t}$ BY DEFINITION OF Y_t			
	Univariate			
	$\dot{x}^*$	$\dot{p}^*$	u^*	Multivariate
M2	120	.049	.000	156
$M2^D$	349	.163	.020	.389

NOTE.—The superscript D designates a Divisia aggregate. $M1^D$, $M3^D$, and L^D are not yet available on the FRB-Penn-MIT quarterly model. Also $\dot{x}^*$ here is the growth rate of nominal GNP, $\dot{p}^*$ is the rate of change in the GNP deflator, and u^* is the total unemployment rate. Sample period, quarterly observations 1969 1–1979 4.

ported to have occurred in the middle of 1974.[11] Over the past decade, interest rates have generally been rising. Hence the opportunity cost (user cost) of holding money has been rising. Under those circumstances, tests of functional stability can be deceptive. Conventional tests of functional stability are most useful when explanatory variable values are replicated or at least lie within the same region both before and after the potential shift period. When explanatory variables are continually moving into new regions, it can be very difficult to separate function shift from specification error.

Suppose, for example, that the true demand for-money function is nonlinear in its variables and parameters but has never shifted. Then the parameters of the best linear approximation will differ over different regions of the space of explanatory variables. If the explanatory variables move into new regions of that space as time passes, then tests of functional stability of a linear approximation are likely to reject stability. Similarly, if the parameters of the function are estimated over one period of time and the estimated function is then used in a dynamic simulation over another time period, the function is likely to track poorly. Fortunately, however, the existence of interest-rate cycles over the past decade has resulted in approximate replication regions that can be explored. In this section we present velocity cross plots permitting investigation of the ability of the economy's true demand-for-money function to replicate. If the true function that generated the data over one time period cannot replicate the data over another time period with similar explanatory variables values, then we can validly conclude that the demand-for-money function has shifted.

---

[11] See Enzler, Johnson, and Paulus (1976); Goldfeld (1976), Tinsley and Garrett, with Friar (1980); and Simpson and Porter (1981). This problem has been most heavily investigated at low levels of monetary aggregation, but it appears to arise also at higher levels of aggregation. This feature of the problem is most troublesome, since it suggests that the "shift" is a result not of explainable substitution within the money market, but rather of a true shift in the economy's transactions technology.

The following exploratory data analysis relates velocity to various commonly used opportunity-cost variables. Three different plotting symbols are used to differentiate between data from three different time periods: (1) data after the purported shift, (2) data from the 5 years preceding the purported shift, and (3) data from the prior decade. We seek to determine whether velocity data acquired from different time periods tend to replicate when the potential explanatory variable retraces the same region during the different time periods. The explanatory variables we consider are the Divisia user cost and Moody's AAA corporate bond rate.

Figures 1 and 2 plot sum M3 against the two potential explanatory variables. Strong evidence exists of a shift in the relationship between velocity and either of the two opportunity-cost measures. Of course the reason could be additional omitted variables. But a stable relationship between velocity and any one of the opportunity-cost variables taken alone does not appear to exist, since the function that generated the data after 1974 does not appear to be able to replicate the earlier data in the middle region of overlapping opportunity-cost values.

Figures 3 and 4 plot Divisia M3 against the same two potential explanatory variables. No clear evidence of function shift remains. For example, with ordinary-least-squares linear regression of Divisia M3 on Moody's AAA corporate bond rate [illegible] of the Chow test of the hypothesis of no para[illegible] .1337. The tail area of the corr[illegible] The result is substantially more [illegible] official aggregate.[12]

Although not reported here, [illegible] 6-month Treasury bill rate and [illegible] Barnett, Spindt, and Offenbacher [illegible] play only M3 plots in this section and in [illegible] the results acquired with M2 and *L* are similar. The available empirical evidence, including the causality tests in the last section, suggests that Divisia *L* is the potentially most interesting aggregate. In addition, Divisia *L* is the aggregate that comes closest to capturing the contributions of all elements of the money market to the economy's monetary service flow. However, updates for Divisia *L* are not currently available with the same frequency as for Divisia M3, and hence Divisia *L* is of less potential policy usefulness than Divisia M3 at present. In any case, the plots for *L* are very similar to those displayed

---

[12] A first-order autoregressive error structure was used.

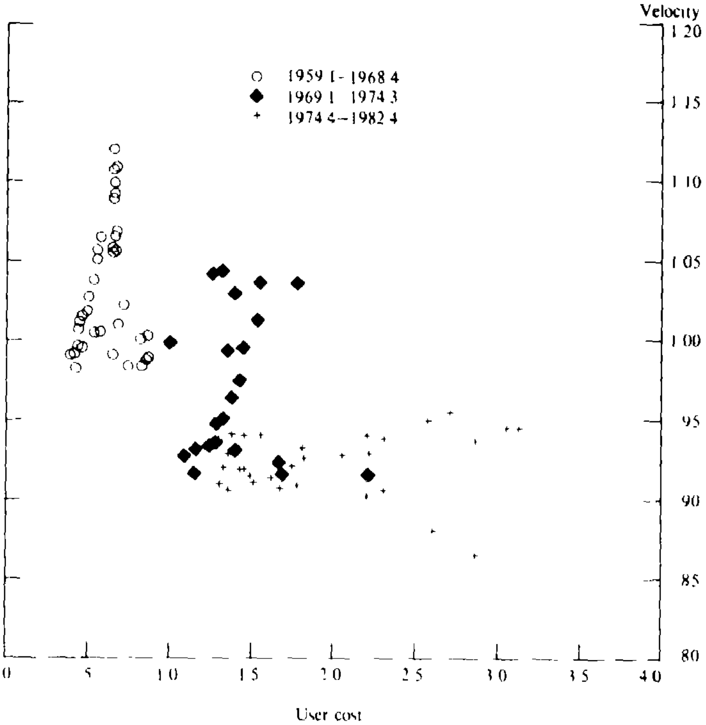


FIG. 1.—Sum M3 velocity versus Divisia user-cost aggregate, quarterly, 1959:1–1982:4.

here for M3, as demonstrated in Barnett, Spindt, and Offenbacher (1981). At the M1 level, as might be expected, the plots for the Divisia aggregate do not differ from those of the sum aggregate by as much as at the higher levels of aggregation.[13]

VI. Money Demand Functions

A. *Specification*

In this section we compare the official monetary aggregates with the Divisia monetary aggregates in terms of the resulting properties of estimated demand-for-money equations. When using the official aggregates, we use the conventional specifications that were used in the literature on demand-for-money function shifts. When using the Divisia aggregates, we use the analogous specifications appropriate to

---

[13] Plots at all levels of aggregation, for both this section and Sec. IX, are available from the authors on request.

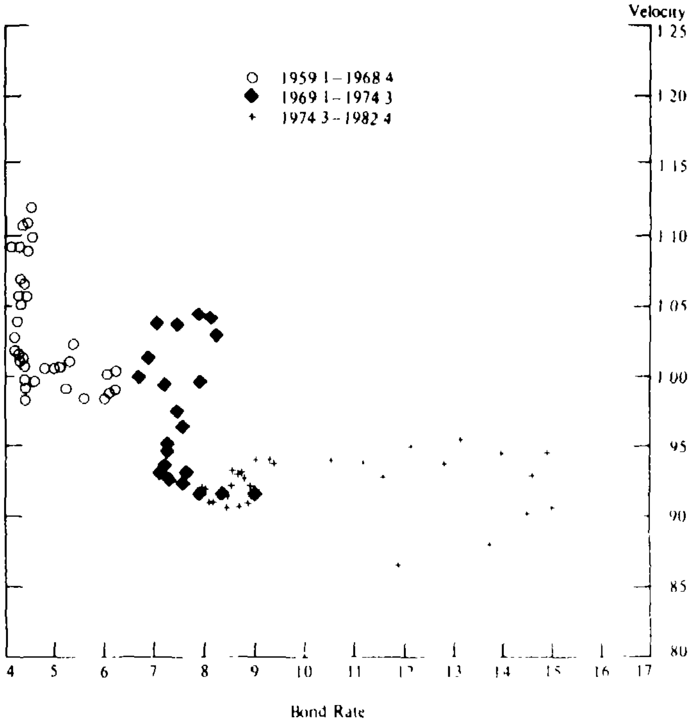


FIG. 2.—Sum M3 velocity versus Moody's [illegible]
1959.1–1982:4.

Divisia aggregation. We exp[illegible]
metric stability, and simulated [illegible]

We use the double log spec[illegible]
literature on "shifting" dema[illegible]

$$\log\left(\frac{M_t}{p_t^*}\right) = \alpha_0 + \alpha_1 \log\left(\quad p_t \right) \quad t$$
$$+ \sum_{k=1}^{K} \alpha_{2+k} \log(OC)_k, \qquad (1)$$

---

[14] All variables are quarterly averages, and all variables except interest rates are seasonally adjusted. Similarly, the Divisia aggregates are constructed from seasonally adjusted quantities and unadjusted interest rates. While this functional form is well known, two aspects of the precise specification require clarification. First, the lagged value of money, which appears as a predetermined variable, is deflated by the current value of the price level to allow for *partial adjustment of nominal balances to price-level shocks* in the short run, while maintaining long-run linear homogeneity of M_t in (Y_t, p_t). Regarding such double log inventory models, see Baumol (1952), Tobin (1956), and Fisher (1978).

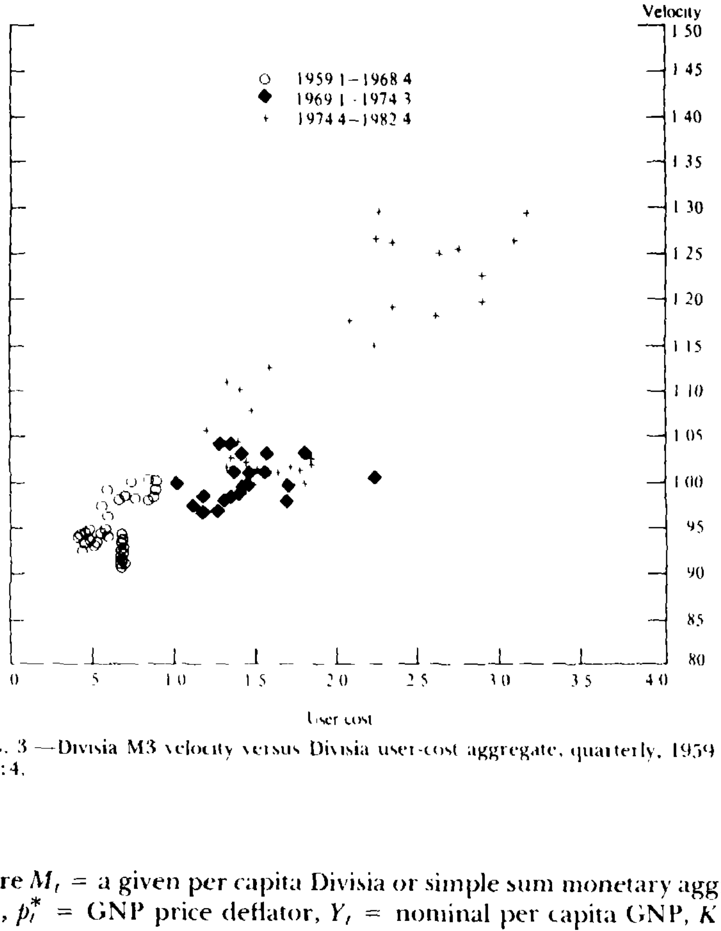


FIG. 3.—Divisia M3 velocity versus Divisia user-cost aggregate, quarterly, 1959:1–1982:4.

where M_t = a given per capita Divisia or simple sum monetary aggregate, p_t^* = GNP price deflator, Y_t = nominal per capita GNP, K = total number of opportunity-cost variables (one or two), and $(OC)_k$ = kth opportunity-cost variable (interest-rate or Divisia user-cost aggregate). The opportunity-cost variables differ among the different equations. When M_t is a summation aggregate, the opportunity-cost variables are interest rates: the commercial paper rate (as a representative market interest rate) and, with M1, the commercial bank passbook rate. When M_t is a Divisia index, all opportunity-cost variables are Divisia user-cost indexes acquired from Divisia price aggregation over the real user costs, π_{it}^*, of equation (3). The Divisia user-cost index is computed for the own price of M_t and, with M1, for a competing aggregated good. The competing opportunity-cost variable is the Divisia user-cost aggregate for those assets that are in L but not in M3.

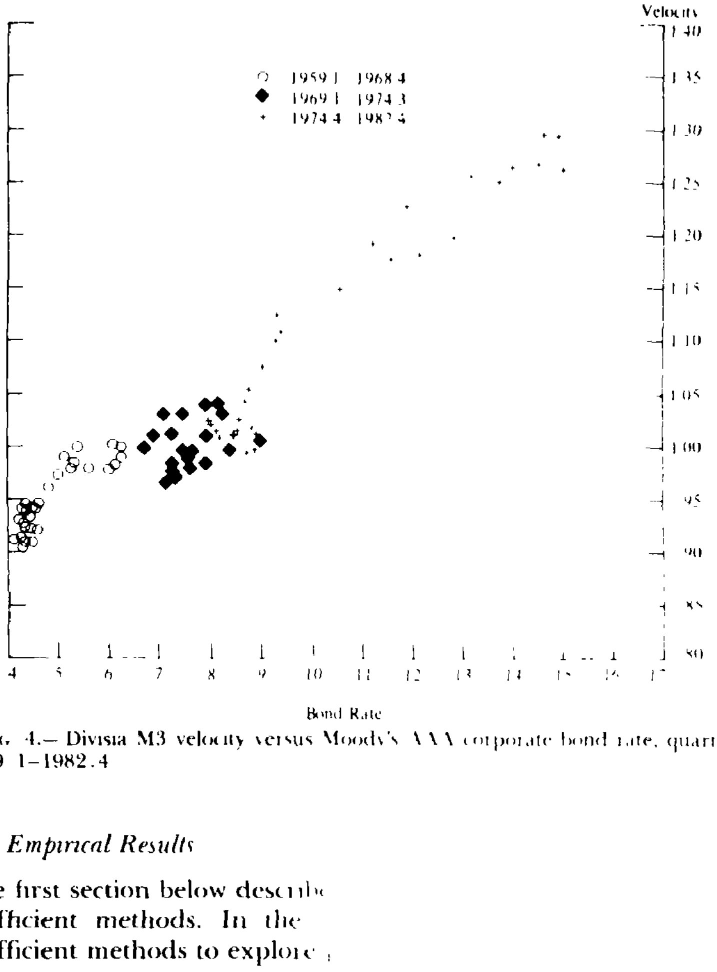


FIG. 4.— Divisia M3 velocity versus Moody's AAA corporate bond rate, quarterly 1959.1–1982.4

B. *Empirical Results*

The first section below describ
coefficient methods. In the
coefficient methods to explore

1. Forecasting

The parameter estimates for equation (4) for both the Divisia and sum aggregates are found in Appendix A.[15] The estimates seem reasonable by conventional standards. Usually the coefficients on real GNP and on the opportunity-cost variables are statistically significant, have the correct sign, and are of reasonable magnitude. Two exceptions are the wrong signs acquired for the coefficient of real GNP with

---

[15] A first-order autoregressive error structure was used.

Divisia M2 and Divisia *L* during 1974:3–1982:4.[16] No such exceptions occurred during 1959:3–1974:2. In the literature on money demand, the most troublesome parameter to interpret has always been the coefficient on the lagged dependent variable (LDV), which has almost always implied unreasonably slow speeds of adjustment. While Divisia aggregation has not totally resolved this problem, the coefficient on the LDV is lower for the Divisia aggregates than for their official sum counterparts in eight of the 12 cases. The eight cases include all four levels of aggregation during the 1959:3–1974:2 sample period. The coefficient on the LDV exceeds 1.0 in three cases, all using the official sum aggregates. With both the Divisia and sum aggregates, the coefficient estimates were most plausible during the 1959:3–1974:2 sample period.

The predictive performance of each equation with its parameters estimated using the 1959:3–1974:2 sample period is summarized by the results that appear in the first two columns of table 3. The first of the columns reports the root mean square error (RMSE) for the predicted growth rates implied by each estimated equation over the period 1974:3–1980:4. The second column reports the mean error of the same growth-rate predictions.[17] The simulation results uniformly favor the Divisia aggregates. At all levels of aggregation, both the RMSE and the absolute value of the mean error for the Divisia forecasts are lower than for their sum counterparts.

2. Parametric Stability

In order to explore parametric stability, we estimate equation (4) with stochastically varying coefficients. We permit the coefficients to be stationary stochastic processes and use the Swamy and Tinsley (1980) asymptotically efficient estimation procedure. With this technique, it is necessary to use the same specification for equation (4) with both the Divisia and official sum aggregates, in order to assure comparability of the coefficient time paths (and test statistics) between results with both sets of aggregates.[18] To avoid prejudicing the results in

---

[16] Also note that the own-price elasticity of the demand for Divisia M3 has the wrong sign during 1959:3–1982:4, and the coefficient of $(OC)_2$ is statistically insignificant for both Divisia and sum M1.

[17] Growth-rate forecast errors are obtained by simulating each equation dynamically. Starting in 1974:3, we obtain predicted levels of the relevant aggregate and then compute the predicted growth rates. The estimation sample period was ended and the forecasts begun in mid-1974, since the widely reported shift in the demand for money function is professed to have occurred at that time.

[18] Otherwise it would be impossible to separate the effect of the different aggregation procedure from the effect of the different equation specification. In fact, if different specifications were used, there would be no formal procedure for determining which coefficient path from one equation to compare with any given coefficient path from the other equation.

TABLE 3

FORECASTING PROPERTIES OF FIXED COEFFICIENT DEMAND FUNCTIONS

MONETARY AGGREGATE	FORECASTING* RMSE†	Mean Error
M1	5 739	− 319
$M1^D$	3.924	− 058
M2	5 141	2.678
$M2^D$	4 966	- 213
M3	5 296	3 015
$M3^D$	4 472	- 150
L	5 617	3.493
L^D	4 109	- 085

NOTE —All data are quarterly The results use parameters estimated for the 1959 3 1974 2 sample period in forecasting from 1974 3 through 1982 4 The *D* superscript on the monetary aggregates designates Divisia aggregates, the others are sums

* Based on growth-rate forecast errors in percentage points per year

† Root mean squared errors of forecasts

favor of the Divisia aggregates, we do not use Divisia user-cost aggregates as explanatory variables, but rather the conventional demand-for-money equation adopted with the official sum M1 aggregate in the last section. In this section, we use that one specification at all levels of aggregation and with both the Divisia and sum aggregates.

The coefficients of equation (4) now are viewed as stochastic processes and hence are written with time subscripts as $\boldsymbol{\alpha}_t = (\alpha_{0t}, \alpha_{1t}, \alpha_{2t}, \alpha_{3t}, \alpha_{4t})$. In accordance with the Swamy and Tinsley (1980) procedure, we further specify that $\boldsymbol{\alpha}_t = \bar{\boldsymbol{\alpha}}$ random vector with mean zer is a vector of parameters (t matrix of parameters.[19] The mean coefficient vector, $\bar{\boldsymbol{\alpha}}$, ar table 4. The estimates appea aggregation with both the sum common problem of slow speed of

In figures 5, 6, and 7, we have plotted for M α_{3t}, α_{4t}), which are the coefficients of the three exogenous variables.[20] Each of the three coefficient paths is substantially more stable with the Divisia aggregate than with the sum aggregate. Furthermore, the cyclical drift evident in the results with the sum aggregate is absent from the Divisia results.

---

[19] Estimates of Φ and Δ are reported in an earlier draft of this paper, which is available from the authors on request.

[20] The coefficient path realizations charted in figs. 5 7 were computed as $\boldsymbol{\alpha}_t = \hat{\boldsymbol{\alpha}} + \mathbf{e}_t$, following the procedures outlined by Swamy and Tinsley (1980) The estimator $\hat{\boldsymbol{\alpha}}$ is consistent for the mean of the stochastic process $\boldsymbol{\alpha}_t$. Our predictions, $\mathbf{e}_t$, of the *coefficient innovation process are based on the minimum norm generalized inverse* solution given by Swamy and Tinsley (1980, p 116, eq [4 10])

TABLE 4

MEAN PARAMETER ESTIMATES AND STABILITY TESTS

	MEAN PARAMETER ESTIMATES							
MONETARY AGGREGATE	Intercept ($\bar{\alpha}_0$)	Lagged Money ($\bar{\alpha}_1$)	Income ($\bar{\alpha}_2$)	Commercial Paper ($\bar{\alpha}_3$)	Passbook Rate ($\bar{\alpha}_4$)	*F*-STATISTIC*	CHOW†	*Df*‡
M1	. .		. .	. .	.	. . .	3.89	6,74
							[.002]	
M1^D	.	. . .	. .	. . .	.	. .	3.86	6,74
							[.002]	
M2	.008	.860	.212	−.032	.020	7.348	2.36	5,76
	(2.12)	(11.46)	(3.21)	(−4.67)	(1.66)	[.000]	[.048]	
M2^D	.008	.896	.135	−.037	.002	3.085	2.32	5,76
	(1.59)	(7.36)	(1.62)	(−2.00)	(.08)	[.000]	[.051]	
M3	.008	.844	.222	−.015	.033	47.449	1.82	5,76
	(.19)	(6.56)	(1.94)	(−.36)	(.20)	[.000]	[.119]	
M3^D	.005	.868	.167	−.035	.003	1.418	1.40	5,76
	(1.61)	(10.24)	(2.98)	(−3.95)	(.21)	[.137]	[.234]	
L	.010	.901	.150	−.014	.012	8.859	1.62	5,76
	(3.92)	(24.99)	(3.75)	(−3.45)	(1.15)	[.000]	[.164]	
L^D	.002	.797	.128	−.023	.020	1.938	.95	5,76
	(.17)	(7.92)	(2.55)	(−3.22)	(.88)	[.019]	[.452]	

NOTE.—Sample period, quarterly data 1959:3–1980:4. Numbers in parentheses are *t*-ratios. Numbers in brackets are tail areas of tests. The *D* superscripts after the monetary aggregates designate Divisia aggregates; the others are sums. The Swamy and Tinsley estimation algorithm failed to converge with the M1 data, so no results could be presented for those data.

* *Df* = 24.62

† The Chow test statistic, for a shift at the end of 1974:2, is an *F*-statistic.

‡ *Df* for the Chow test *F*-statistic.

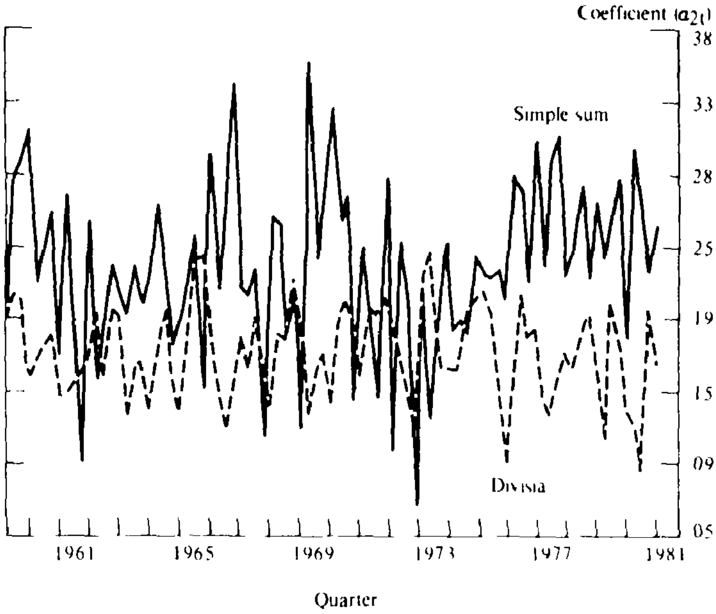


FIG. 5.—Time path of the income coefficient, α_{2t}, 1959:2–1980:4

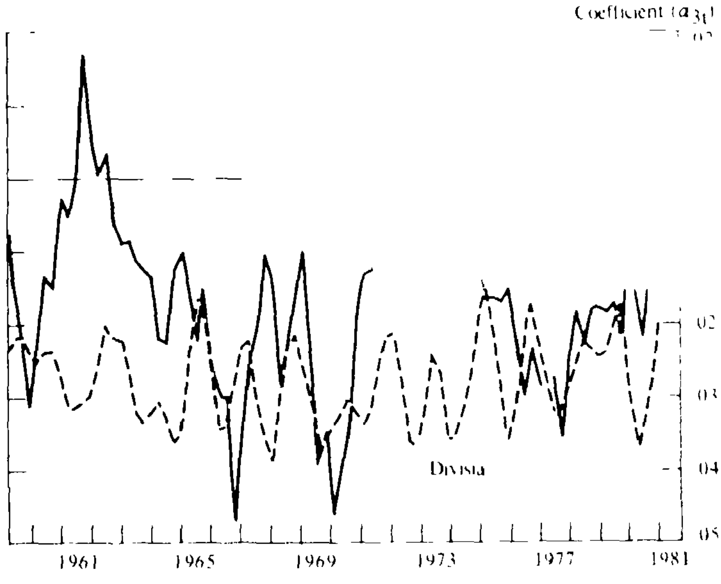


FIG. 6.—Time path of commercial paper rate coefficient, α_{3t}, 1959:2–1980:4

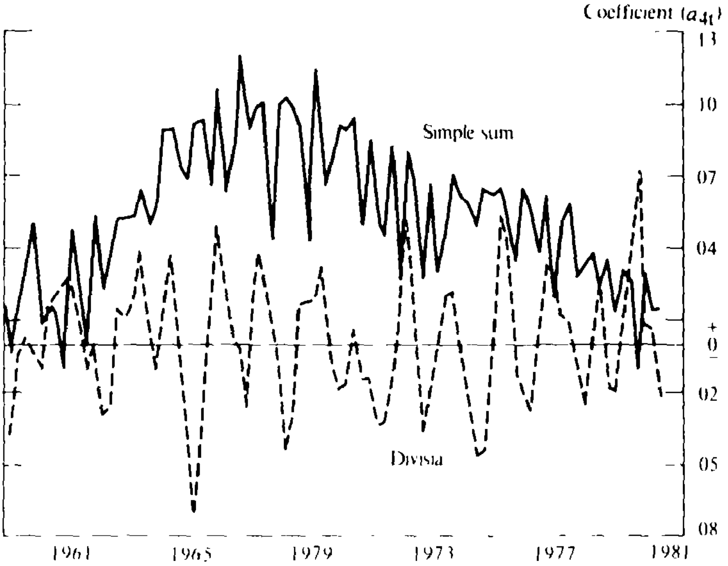


FIG. 7.—Time path of passbook rate coefficient, α_{4t}, 1959.2 -1980.4

The observations above on coefficient path behavior are formally verified by the *F*-statistics in the sixth column of table 4. Those *F*-statistics are for the test of the hypothesis that equation (4) has constant coefficients and an additive first-order autoregressive error structure. At every level of aggregation, the *F*-statistic with the Divisia aggregate is lower than with the simple sum aggregate. Since the tail area of the test is inversely related to the level of the *F*-statistic, the hypothesis of parametric stability was more acceptable with the Divisia than with the sum aggregates. The result is most striking at the M3 level. At the .05 level of significance, stability of the demand function for the sum aggregate would be decisively rejected, while stability would be decisively accepted for the corresponding Divisia aggregate. Nevertheless, at all other levels of aggregation, stability would be rejected at the .05 significance level with either the sum or Divisia aggregate.

The seventh column of table 4 contains the *F*-statistic for a Chow test of the hypothesis of no break in regimes after 1974:2. Since we do not produce stochastic coefficient paths in this test, the problem of comparability with different specifications is less severe than with the stochastic coefficient results.[21] As a result, we use the same equation specifications as in the last section for each aggregate. The F-statistic is larger for the Divisia aggregate than for the sum aggregate at each level of aggregation. These results are uniformly favorable to the Divisia aggregates. However, at a fixed .05 level of significance, the hypothesis of no shift would be accepted for both the Divisia and sum aggregates at the M3 and L levels of aggregation. The hypothesis would be rejected for both Divisia and sum M1. The hypothesis would be marginally accepted for Divisia M2 and marginally rejected for sum M2.

---

[21] In addition, using different specifications decreases the degree of partial redundancy of this test with the stochastic coefficients *F*-test.

Searching over tables 3 and 4 for a uniformly best aggregate, we find none. In terms of forecasting, Divisia M1 and Divisia L were most successful. In terms of stability, Divisia L was most successful with the Chow test, but Divisia M3 was most successful with the stochastic coefficient F-test.

VII. Reduced-Form Equations

A. The Equations

Another basis for comparing various monetary aggregates is a comparison of their performance in reduced-form equations.[22] These equations relate the growth rate of GNP to current and lagged money growth rates and to current and lagged values of a fiscal policy variable. Such an equation is interpreted as a reduced-form equation from an unspecified structural econometric model.

The specification adopted here is from Carlson (1980) and has the form

$$\dot{Y}_t = \beta_0 + \sum_{i=1}^{11} \beta_i$$

where $\dot{Y}_t$ = annualized percent
annualized percentage rate of gr
tary aggregate during quarter t,
of growth of high employment federal
are estimated both with and without the constraint, [illegible]
which is necessary for steady-state (long-run) superneutrality.[23]

B. Results

Table 5 summarizes the results. The criteria for comparing the findings are essentially the same as for the money demand equations

---

[22] The interpretation of such reduced-form equations is subject to a number of well-known difficulties (see, e.g., Lucas 1976).

[23] Strictly speaking, superneutrality refers to the lack of any effect of inflation on the level of real output; here the term refers to the lack of any effect of inflation on real output growth. The distributed lags are third-order Almon polynomial distributed lags. All the equations were estimated by the Cochrane-Orcutt technique.

TABLE 5

FIT, FORECASTING, AND STABILITY PROPERTIES OF ESTIMATED REDUCED-FORM EQUATIONS

SPECIFICATION	MONETARY AGGREGATE	R^2	FORECASTING*		STABILITY	
			RMSE†	Mean Error	Chow‡	LR§
Unconstrained	M1	2872	4.795	− 092	96 [.466]	7.72 [.461]
Constrained	M1	2937	4 813	−.267	1.15 [.341]	8.89 [.261]
Unconstrained	$M1^b$	2527	3 417	− 276	1.17 [.326]	9 03 [.251]
Constrained	$M1^b$	.2528	3 527	−.710	1.46 [194]	10.14 [181]
Unconstrained	M2	2458	4 829	− 222	2.08 [.048]	16.38 [.037]
Constrained	M2	.2531	4.829	−.233	.69 [682]	4.83 [.680]
Unconstrained	$M2^b$	3155	4 811	954	47 [871]	4 49 [.810]

Constrained	M2^D	3154	4 845	1 044	.53 [.813]	4.09 [.769]
Unconstrained	M3	2111	5.032	396	1.82 [086]	13.89 [.080]
Constrained	M3	.2171	5 024	− 206	1.10 [.370]	8.83 [.265]
Unconstrained	M3^D	.3002	4 833	1.335	.58 [.789]	4.84 [775]
Constrained	M3^D	3019	3 028	1 239	.43 [.880]	4 66 [.701]
Unconstrained	L	.2178	5.540	−2 033	.84 [.557]	8.19 [.415]
Constrained	L	2152	5 412	−1 699	1.06 [.394]	8.79 [.357]
Unconstrained	L^D	.3376	2 814	− 282	.77 [.627]	6 38 [603]
Constrained	L^D	[illegible]	4 635	301	52 [.818]	4.49 [.722]

NOTE —Sample period quarterly data al areas of tests The D superscript on the monetary aggregates designates Divisia aggregates the others are sums The stability results test for a shift at se parameters estimated for the 1959 3–1974 2 sample period in forecasting from 1974 3 through 1982 4.

* Based on growth-rate forecasting err

† Root mean squared errors of forecasts

‡ The Chow test statistic is an F-statistic unconstrained test are 8 78

§ The LR test statistic is the asymptotic s a limiting χ^2 distribution The number of df is the same as the numerator df for the Chow F-statistic forecasting performance and stability tests. We again split the sample after 1974:2 in acquiring the forecasting and stability results. Further details for the estimation results can be found in Barnett, Spindt, and Offenbacher (1981). Table 5 contains all of the results on forecasting performance and stability behavior.

The results in table 5 are mixed, although certain patterns are evident. The performance of the Divisia aggregates judged relative to the corresponding sum aggregates gradually improves as the level of aggregation increases. At the highest level of aggregation, Divisia *L* outperforms sum *L* relative to all of the criteria in table 5. At lower levels of aggregation, the forecasting results depend heavily on the criterion used. Relative to RMSE, the Divisia aggregates usually outperform the sum aggregates, but the reverse conclusion is acquired relative to mean error. In terms of both stability (test tail area) and fit (R^2), the sum aggregate outperforms the Divisia aggregate at the M1 level of aggregation, but that conclusion is reversed relative to both criteria at all other levels of aggregation.

Comparisons across levels of aggregation do not reveal any single best aggregate, although Divisia *L* generally did very well in table 5.

VIII. Divisia Second Moments

The right-hand side of equation (1) is in the form of a statistical expectation or first moment. This result follows from the fact that $\Sigma_j \bar{s}_{jt} = 1$, and $\bar{s}_{jt} \geq 0$, so that each $\bar{s}_{jt}$ can be viewed as a probability from a discrete probability distribution. Hence we can define corresponding Divisia second moments in the obvious manner.[21] In Appendix B we show how those Divisia second moments can be used to complement the Divisia quantity index (or Divisia quantity mean) by providing a dispersion measure of "potential aggregation error." We find that the potential aggregation error of the Divisia monetary aggregates has not been subject to appreciable cyclical variation over our sample period of 1969:1–1982:4, and any slight cyclical variation decreases as the level of aggregation increases.

IX. Controllability

Since this paper primarily explores the relationship between monetary (intermediate) targets and final targets, we do not here extensively investigate controllability, which is defined in terms of the relationship between instruments and intermediate targets. Nevertheless we have explored the stability of multipliers between instruments and intermediate targets. The instruments we considered were the monetary base, total reserves, and nonborrowed reserves. The intermediate targets we considered were Divisia and sum M1, M2, M3, and *L*. All of the multipliers were erratic over time except for the ratio of the Divisia aggregates to the base. Those ratios exhibited stable long-run trends. The slope of the trend decreased as the level of aggregation increased, the ratio becoming approximately constant in the long run for Divisia *L* (with a short-run cycle correlating with interest rates).

---

[21] See Barnett, Spindt, and Offenbacher's (1981) eqq. (12.1)–(12.7) for formal definitions.

Most of these results are available in Barnett and Spindt (1982).[25] Their results with sum and Divisia M3 are updated and displayed below. Using monthly data, we cross plot in figures 8 and 9 the base money multiplier (monetary aggregate divided by monetary base) against Moody's Baa (average quality debt) corporate bond rate for both sum and Divisia M3. As in figures 1–4, we use three plotting symbols for three different time periods.[26] As is evident from figure 9, an interest rate has high explanatory power relative to the base multiplier of Divisia M3. The same cannot be said for sum M3, as is evident from the broad dispersion in figure 8.

Since relevant microeconomic foundations for the supply-of-money function have only recently been developed, theoretical interpretation at this time of the results above would be speculative.[27] One might, for example, view the monetary base as both the output of the Federal Reserve System and a wealth constraint on the private sector.
In this interpretation, a change in the
to re-equilibrate itself and there
tary service flow. Since the bro
sure that service flow, a stable re
at any given level of interest ra
Divisia M3 or *L*. Figure 9 tends

X. Conclusion

Aggregation theory favors the Divisia quantity index over the sum index as a measure of the quantity of an aggregated economic good,

---

[25] For further consideration of the predictability of the base multipliers of the Divisia monetary aggregates, see Spindt (1984).

[26] In order to detrend the series, we measure each variable (whether base multiplier or interest rate) as its residual in a linear regression of the variable on time

[27] Barnett and Hinich (in press) have derived the supply functions for the Divisia monetary aggregates and have estimated those functions through the application of Hilbert transform methods. The results provide a direct test of controllability Also see Hancock (1984) for further relevant theory

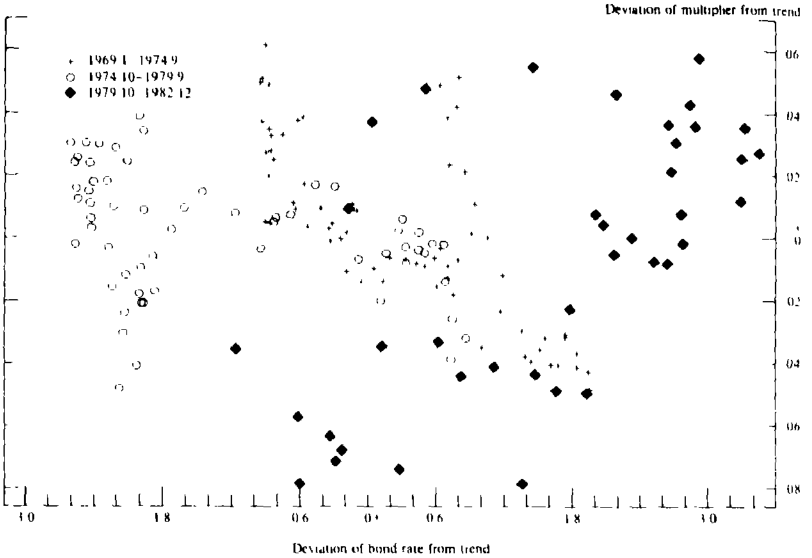


FIG. 8.—Deviation from time trend of sum M3 base money multiplier versus deviation from time trend of Moody's Baa corporate bond rate, monthly, 1969:1–1982:12

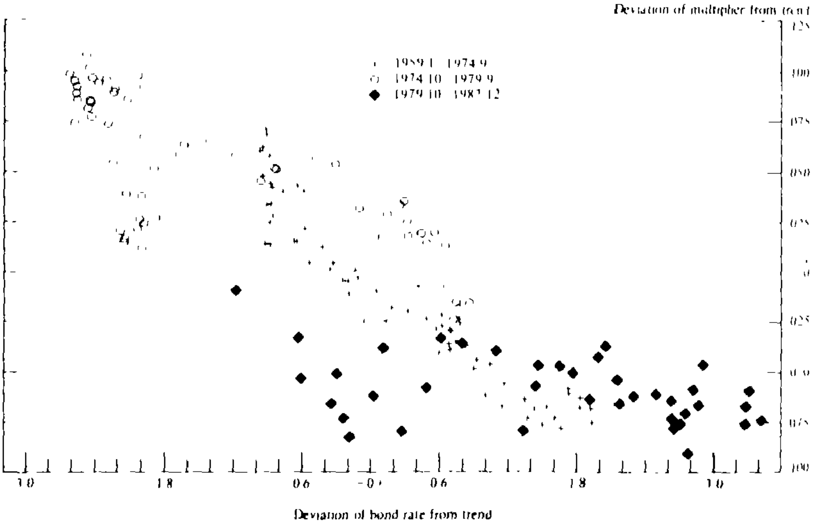


FIG. 9.—Deviation from time trend of Divisia M3 base money multiplier versus deviation from time trend of Moody's Baa corporate bond rate, monthly, 1969:1 1982:12.

when the components are not perfect substitutes. Barnett's (1980*a*, 1981*a*) tests using monetary data resulted in rejection of the necessary and sufficient conditions in aggregation theory for sum aggregation (see also Feige and Pearce 1977; Offenbacher 1979; Ewis and Fisher 1984*a*, 1984*b*; Serletis 1984*a*, 1984*b*). The present paper systematically compares the empirical performance of the Divisia and the sum monetary aggregates relative to various criteria relevant to policymaking. Neither the Divisia aggregates nor the sum aggregates uniformly dominated the others relative to all of the criteria considered, and no one aggregate, whether Divisia or sum, was uniformly best. However, some general tendencies are evident from these results, as can be seen from the following summary of results.

In the causality tests, the Divisia aggregates generally performed better than the corresponding sum aggregates, although sum M2 did rather well. Divisia *L* was perhaps the best aggregate in those tests. In terms of the demand-for-money functions, the best forecasting results were acquired with Divisia M1 and Divisia *L*. The most stable demand-for-money functions were acquired with Divisia M3 and Divisia *L*. In addition, the velocity function for Divisia M3 was found to be stable. In the reduced-form comparisons, sum M1 performed better than Divisia M1, but at higher levels of aggregation the Divisia aggregates became increasingly superior to the corresponding sum aggregates, with Divisia *L* usually providing the best reduced-form results.

In earlier work, using information theory, we found that the Divisia aggregates tend to perform better than the sum aggregates as indicators, especially at high levels of [illegible] with the FRB-Penn-MIT quarte[illegible] result with M2. In addition, us[illegible] find that the Divisia monetar[illegible] variation in potential aggregati[illegible] Divisia *L* was best. In addition, [illegible] supporting our earlier results on [illegible] broad Divisia aggregates.

In short, at the lowest (M1) level of aggregation [illegible] ing results in our comparisons between the sum and the Divisia aggregate. However, at higher levels of aggregation, the Divisia aggregates generally tend to perform better than the sum aggregates, with that degree of superiority tending to increase at increasing levels of aggregation.[28] Since the divergence between the time paths of the Divisia and the sum aggregates increases as the level of aggregation increases (and the discrepancies between the two weighting methods increases), the power of any tests comparing the two aggregation methods should be expected to increase as the aggregation level increases. With so many criteria being considered, the selection of a "best" aggregate is a hazardous matter. Cagan's (1982) results, based on fewer criteria and an earlier sample period, generally favored Divisia M1. While no aggregate was uniformly best relative to all of our criteria, our results reflect most favorably on Divisia *L*.

---

[28] Preliminary results suggest that these conclusions also apply to Canadian data (see Cockerline and Murray 1981).

Appendix A

The results of the error structure analysis for the Sims test are displayed in table A1. The parameter estimates for the fixed coefficients demand-for-money functions are displayed in tables A2–A5.

TABLE A1

TAIL AREAS OF THE TEST THAT THE FIRST EIGHT AUTOCORRELATIONS OF THE SIMS DISTURBANCE TERMS ARE ALL ZERO

Monetary Measure	GNP to Money	Money to GNP
M1	.459	619
$M1^D$	.621	.520
M2	091	.735
$M2^D$	368	.420
M3	127	280
$M3^D$	.483	.652
L	.197	363
L^D	533	.212

NOTE – Sample period, quarterly observations 1959.1 -1982 4 The superscript *D* designates a Divisia aggregate The other aggregates are sums

TABLE A2

MONEY DEMAND ESTIMATES FOR DIVISIA AND SUM M1, DOUBLE LOGARITHMIC SPECIFICATION

Aggregate and Period of Fit	Constant	Real GNP	*OC*1*	*OC*2†	Lagged Dependent	ρ	SEE	R^2	D-W	Box-Pierce‡
Divisia M1·										
1959 3–1974:2	– 1 243	.1065	– 0070	– 0022	7919	.538	.0054	976	1.689	13.321
	(437)	(0216)	(0067)	(.0011)	(0796)	(.1088)				
1974 3–1982 4	– 953	[illegible]	[illegible]	– 0013	8177	– 2150	0062	956	2 071	7 835
	(360)			(0028)	(.0734)	(.1675)				
1959 3–1982:4	– 255			0000	9691	.2458	.0065	967	1 895	10.929
	(148)			(0011)	(0274)	(.1000)				
Sum M1										
1959.3–1974·2	– 1 059			– 0219	8165	4364	.0042	909	1.900	11.141
	(440)			(0139)	(0841)	(1162)				
1974 3–1982 4	– 401			– 0930	.9170	– 3557	.0018	970	2 140	11 138
	(214)			(1167)	(0674)	(.1603)				
1959 3–1982 4	– 059			0000	1 0018	– 0427	.0063	990	1.957	8.124
	(073			(0089)	(0148)	(1031)				

NOTE — Standard errors are in parenthe

* Divisia *OC*1 = Divisia user cost for N [illegible]

† Divisia *OC*2 = Divisia user cost for [illegible] state *OC*2 = passbook rate

‡ Box Pierce χ^2 statistic for first 12 sa

TABLE A3

Money Demand Estimates for Divisia and Sum M2, Double Logarithmic Specification

Aggregate and Period of Fit	Constant	Real GNP	*OC**	Lagged Dependent	ρ	SEE	R^2	D-W	Box-Pierce†
Divisia M2:									
1959:3–1974:2	−1.006	.1421	−.0102	.8484	.7110	.0053	.996	1.656	10.983
	(.537)	(.0570)	(.0050)	(.0870)	(.0908)				
1974:3–1982:4	−.119	−.1019	−.0248	.9347	.1173	.0095	.987	1.984	6.222
	(.231)	(.0381)	(.0111)	(.0416)	(.1703)				
1959:3–1982:4	−.098	.0409	−.0088	.9916	.5877	.0080	.992	2.034	5.508
	(.220)	(.0281)	(.0054)	(.0357)	(.0835)				
Sum M2:									
1959:3–1974:2	−.383	.1409	−.0304	.9573	.4133	.0039	.999	1.935	7.927
	(.300)	(.0498)	(.0040)	(.0428)	(.1176)				
1974:3–1982:4	−1.014	.1705	−.0170	.8479	−.1874	.0054	.985	1.960	4.294
	(.455)	(.0696)	(.0049)	(.0687)	(.1685)				
1959:3–1982:4	−.721	.1764	−.0238	.9057	.4644	.0046	.999	2.048	4.910
	(.275)	(.0455)	(.0035)	(.0394)	(.0913)				

Note.—Standard errors are in parentheses.

* Divisia *OC* = Divisia user cost for M2; sum *OC* = commercial paper rate.

† Box-Pierce χ^2 statistic for first 12 sample autocorrelations.

TABLE A4

MONEY DEMAND ESTIMATES FOR DIVISIA AND SUM M3, DOUBLE LOGARITHMIC SPECIFICATION

Aggregate and Period of Fit	Constant	Real GNP	*OC**	Lagged Dependent	ρ	SEE	R^2	D-W	Box-Pierce†
Divisia M3,									
1959.3–1974 2	– 855	1499	– 0129	.8799	7615	0056	997	1.646	12.039
	(522)	[illegible]	(0055)	(0817)	(.0837)				
1974 3–1982 4	– 196		0219	9319	– 2296	0086	985	1.998	5 933
	(257)		(0108)	(0484)	(1669)				
1959 3–1982 4	– 098		0088	9916	5876	.0080	.992	2 035	5 019
	(220)		(0153)	(.0357)	(.0835)				
Sum M3:									
1959 3–1974 2	– 02[illegible]		[illegible]57	1 0086	5951	0049	.999	1.749	10.947
	(382		[illegible]51)	(0519)	(1038)				
1974 3–1982 4	– 349		[illegible]57	9588	2671	0041	994	1.943	6 898
	(28[illegible]		[illegible]32)	(0446)	(1653)				
1959 3–1982 4	– 50[illegible]		[illegible]96	9394	6566	0048	999	1.863	7 002
	(26		[illegible]39)	(0366)	(0777)				

NOTE — Standard errors are in paren

* Divisia *OC* = Divisia user cost for M

† Box Pierce χ^2 statistic for first 12 sa

TABLE A5

MONEY DEMAND ESTIMATES FOR DIVISIA AND SUM L, DOUBLE LOGARITHMIC SPECIFICATION

Aggregate and Period of Fit	Constant	Real GNP	*OC**	Lagged Dependent	ρ	SEE	R^2	D-W	Box-Pierce†
Divisia L:									
1959:3–1974:2	-1.152	.1502	-.0100	.8229	6820	0046	.996	1.739	5.081
	(.519)	(.0502)	(.0043)	(0852)	(.0944)				
1974·3–1982:4	-.131	- 0750	- 0187	.9426	.3756	0079	.988	2.082	4.912
	(.258)	(.0417)	(0099)	(0481)	(.1589)				
1959:3–1982:4	-.200	.0501	-.0085	9749	.6729	.0063	.993	2.059	5.013
	(.231)	(.0252)	(.0045)	(.0386)	(0764)				
Sum L:									
1959:3–1974:2	.156	.0442	-.0184	1.0345	.4380	.0036	.999	1.968	6.789
	(.338)	(.0519)	(0036)	(0490)	(.1160)				
1974:3–1982:4	-.463	.1240	- 0065	.9450	-.3530	.0034	.997	1.888	11.676
	(.210)	(0363)	(.0029)	(.0310)	(.1605)				
1959:3–1982:4	-.456	.1238	-.0110	.9442	.5607	0037	.999	1.995	3.090
	(191)	(.0321)	(0027)	(.0270)	(.0854)				

NOTE.—Standard errors are in parentheses

* Divisia *OC* = Divisia user cost for L; sum *OC* = commercial paper rate

† Box-Pierce χ^2 statistic for first 12 sample autocorrelations

Appendix B

Let K_{it} be the Divisia quantity variance, and let J_{it} be the Divisia user-cost price variance. Also let Γ_{it} be the Divisia price-quantity covariance, and let Ψ_{it} be the Divisia variance of the shares, $\mathbf{s}_t$. Theil (1967, p. 155) has shown that the Divisia second moments are related by the following equality:

$$K_{it} = \Psi_{it} - J_{it} - 2\Gamma_{it}. \tag{B1}$$

Equation (B1) permits us to decompose K_{it}, which is the Divisia second moment of primary interest in monetary policy.

If the aggregation conditions described in Section III are exactly satisfied, then the continuous-time Divisia index always exactly equals the economic aggregate, and then the aggregation error of the Divisia index, (1), is very small. However, if the aggregation conditions from Section III are only approximately satisfied, the aggregation error (difference from the exact economic aggregate) of the index (1) could be nonnegligible. The Divisia quantity variance, K_{it} (or the corresponding coefficient of variation), can be used as a measure of potential aggregation error. That interpretation of K_{it} can be seen from the fact that K_{it} measures the dispersion of growth rates between the components of the Divisia monetary quantity index. Clearly, if $K_{it} = 0$, then any index number or aggregator function that is linearly homogeneous in the components is as good as any other, since all aggregates grow at the common rate at which each component grows. As K_{it} increases, the quality of the index number formula becomes increasingly important and the risks of aggregation error and information loss increase.

The Divisia share variance, Ψ_{it}, is a measure of the change in the dispersion of the Divisia weights. We might expect that as interest rates increase, Ψ_{it} would rise, since relative prices (user costs) between rate-regulated and rate-unregulated monetary assets move away from 1.0. It might then further be thought that the Divisia quantity variance, K_{it}, would also increase with increasing interest rates, as a result of the increasing dispersion of its weights. However, (B1) shows that this conclusion
in J_{it} resulting from the increasing
rates could offset the increasing

To explore this possibility, we
their correlation with the interest
lation coefficients between the D
table B1. Neither K_{it} nor Γ_{it} cor
both Ψ_{it} and J_{it} do, with nearly
pothesized in the previous parag

TABLE B1

CORRELATIONS OF DIVISIA SECOND MOMENTS WITH THE FEDERAL FUNDS RATE

MONETARY AGGREGATE	DIVISIA SECOND MOMENT			
	K	Ψ	J	Γ
M2	19	61	49	10
M3	13	39	17	08
L	10	37	36	00

NOTE.—Sample period, quarterly observations 1969 1-1982 4 ness cycle tend to cancel each other out continuously, so that K_{it} remains largely independent of the business and interest rate cycles.

Hence our degree of confidence in the quality of the Divisia monetary quantity index, equation (1), should not be altered by variations in interest rates. Variations in the potential aggregation error, K_{it}, of the Divisia monetary aggregates bear little relationship to the business cycle. In addition, the correlation decreases as the level of aggregation increases.

References

Barnett, William A. "The User Cost of Money." *Econ. Letters* 1, no. 2 (1978): 145–49.

———. "Economic Monetary Aggregates: An Application of Index Number and Aggregation Theory." *J. Econometrics* 14 (September 1980): 11–48. (*a*)

———. "Economic Monetary Aggregates: Reply." *J. Econometrics* 14 (September 1980): 57–59. (*b*)

———. *Consumer Demand and Labor Supply: Goods, Monetary Assets, and Time.* Amsterdam: North-Holland, 1981. (*a*)

———. "The New Monetary Aggregates: A Comment." *J. Money, Credit, and Banking* 13 (November 1981): 485–89. (*b*)

———. "Divisia Indices." In *Encyclopedia of Statistical Sciences*, edited by Samuel Kotz and Norman L. Johnson. Vol. 2. New York: Wiley, 1982. (*a*)

———. "The Optimal Level of Monetary Aggregation." *J. Money, Credit, and Banking* 14, no. 4, pt. 2 (November 1982): 687–710. (*b*)

———. "New Indices of Money Supply and the Flexible Laurent Demand System." *J. Bus. and Econ. Statis.* 1 (January 1983): 7–23. (*a*)

———. "Understanding the New Divisia Monetary Aggregates." *Rev. Public Data Use* 11 (December 1983): 349–55. (*b*)

———. "Developments in Monetary Aggregation Theory." *Econometric Rev.* (1984), in press. (*a*)

———. "Recent Monetary Policy and the Divisia Monetary Aggregates." *American Statis.* 38 (August 1984): 165–72. (*b*)

Barnett, William A., and Hinich, Melvin J. "Derivation and Estimation of the Supply Functions for the Divisia Monetary Aggregates." In *Proceedings of the Second Austin Symposium in Economics. Conference on New Approaches to Monetary Policy*, edited by William A. Barnett and Kenneth J. Singleton. Cambridge: Cambridge Univ. Press, in press.

Barnett, William A.; Offenbacher, Edward K.; and Spindt, Paul A. "New Concepts of Aggregated Money." *J. Finance* 36 (May 1981): 497–505.

Barnett, William A., and Spindt, Paul A. "The Velocity Behavior and Information Content of Divisia Monetary Aggregates." *Econ. Letters* 4, no. 1 (1979): 51–57.

———. "Divisia Monetary Aggregates: Their Compilation, Data, and Historical Behavior." Fed. Reserve Bd. Staff Study no. 116. Washington: Publications Services, Fed. Reserve Bd., 1982.

Barnett, William A.; Spindt, Paul A.; and Offenbacher, Edward K. "Empirical Comparisons of Divisia and Simple Sum Monetary Aggregates." Conference Paper no. 122. Cambridge, Mass.: N.B.E.R., 1981.

Barro, Robert J., and Sahasakul, Chaipat. "Measuring the Average Marginal Tax Rate from the Individual Income Tax." *J. Bus.* 56 (October 1983): 419–52.

Barth, James R.; Kraft, Arthur; and Kraft, John. "The 'Moneyness' of Financial Assets." *Applied Econ.* 9 (March 1977): 51–61.

Baumol, William J. "The Transactions Demand for Cash: An Inventory Theoretic Approach." *Q.J.E.* 66 (November 1952): 545–56.

Bisignano, J. "Real Money Substitutes." Working Paper no. 17. San Francisco: Fed. Reserve Bank, 1974.

Cagan, Phillip. "The Choice among Monetary Aggregates as Targets and Guides for Monetary Policy." *J. Money, Credit, and Banking* 14, no. 4, pt. 2 (November 1982): 661–86.

Carlson, Keith M. "Money, Inflation, and Economic Growth: Some Updated Reduced Form Results and Their Implications." *Fed. Reserve Bank St. Louis Rev.* 62 (April 1980): 13–19.

Caves, Douglas W.; Christensen, Laurits R.; and Diewert, W. Erwin. "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity." *Econometrica* 50 (November 1982): 1393–1414. (*a*)

———. "Multilateral Comparisons of Output, Input, and Productivity Using Superlative Index Numbers." *Econ. J* 92 (March 1982): 73–86. (*b*)

Chetty, V Karuppan. "On Measuring the Nearness of the Near-Moneys." *A.E.R.* 59 (June 1969): 270–81.

———. "On Measuring the Nearness of Near-Moneys: Reply." *A.E R.* 62 (March 1972): 226–29.

Cockerline, Jon, and Murray, John. "Superlative Monetary Aggregation: Some Preliminary Results." Mimeographed. File no. 105-7-4. Ottawa. Bank of Canada, 1981.

Diewert, W. Erwin. "Exact and Superlative Index Numbers." *J. Econometrics* 4 (May 1976): 115–45.

———. "Aggregation Problems in the Measurement of Capital. In *The Measurement of Capital*, edited by Dan Usher. Chicago: Univ. Chicago Press (for N.B.E.R), 1980.

———. "The Economic Theory of Index Numbers. A Survey " In *Essays in the Theory and Measurement of Consumer Behaviour in Honour of Sir Richard Stone*, edited by Angus Deaton. Cambridge: Cambridge Univ. Press, 1981

Donovan, Donal J. "Modeling the Demand for Liquid Assets An Application to Canada." *Internat Monetary Fund Staff Papers* 25 (December 1978) 676–704.

Dorfman, John R. "What Explosi

Enzler, Jared J ; Johnson, Lewis
Money Demand." *Brookings P*
80.

Ewis, Nabil A., and Fisher, Dou
Substitutability between Mone
Fourier Flexible Form " *J. Mac*

———. "The Translog Utility Funct
United States." *J Money, Credit, and Ba*

Feige, Edgar L., and Pearce, Douglas K " The Substitutability [illegible] Near-Monies: A Survey of the Time-Series Evidence." *J Econ Literature* 15 (June 1977): 439–69

———. "The Casual Causal Relationship between Money and Income Some Caveats for Time Series Analysis " *Rev Econ and Statis* 61 (November 1979): 521–33.

Fisher, Douglas. *Monetary Theory and the Demand for Money* New York: Wiley. 1978.

Fisher, Irving. *The Making of Index Numbers A Study of Their Varieties, Tests, and Reliability*. Boston: Houghton Mifflin, 1922

Friedman, Milton, and Schwartz, Anna J *Monetary Statistics of the United States*

Estimates, Sources, Methods. New York: Columbia Univ. Press (for N.B.E.R.), 1970.

Goldfeld, Stephen M. "The Case of the Missing Money." *Brookings Papers Econ. Activity,* no. 3 (1976), pp. 683–730.

Granger, Clive W. J., and Newbold, Paul. *Forecasting Economic Time Series.* New York: Academic Press, 1977.

Hancock, Diana. "The Financial Firm: Production with Monetary and Nonmonetary Goods." Mimeographed. Vancouver: Univ. British Columbia, 1984.

Haugh, Larry. "The Identification of Time-Series Interrelationships with Special Reference to Dynamic Regression." Ph.D. dissertation, Univ. Wisconsin—Madison, 1972.

Hawtrey, Ralph G. "Money and Index-Numbers." *J. Royal Statis Soc.* 93, pt. 1 (1930): 64–85. Reprinted in Hawtrey, Ralph G. *The Art of Central Banking.* London: Longmans Green, 1932. Also reprinted in *Readings in Monetary Theory,* edited by Friedrich A. Lutz and Lloyd W. Mints. Homewood, Ill.: Irwin (for American Econ. Assoc.), 1951.

Klein, Benjamin. "Competitive Interest Payments on Bank Deposits and the Long-Run Demand for Money." *A.E.R.* 64 (December 1974): 931–49.

Lucas, Robert E., Jr. "Econometric Policy Evaluation: A Critique." In *The Phillips Curve and Labor Markets,* edited by Karl Brunner and Allan H. Meltzer. Amsterdam: North-Holland, 1976.

Moroney, John R., and Wilbratte, Barry J. "Money and Money Substitutes: A Time Series Analysis of Household Portfolios." *J. Money, Credit, and Banking* 8 (May 1976): 181–98.

Offenbacher, Edward K. "The Substitutability of Monetary Assets." Mimeographed. Washington: Bd. Governors Fed. Reserve System, 1979.

———. "Economic Monetary Aggregates: Comment." *J. Econometrics* 14 (September 1980): 55–56.

Pierce, David A., and Haugh, Larry D. "Causality in Temporal Systems: Characterizations and a Survey." *J. Econometrics* 5 (May 1977): 265–93.

Pudney, Stephen E. "An Empirical Method of Approximating the Separable Structure of Consumer Preferences." *Rev. Econ. Studies* 48 (October 1981): 561–77.

Schwert, G. William. "Tests of Causality: The Message in the Innovations." In *Three Aspects of Policy and Policymaking. Knowledge, Data and Institutions,* edited by Karl Brunner and Allan H. Meltzer. Amsterdam. North-Holland, 1979.

Serletis, Apostolos "The Substitutability and Separability of Monetary Assets." Ph.D. dissertation. Hamilton, Ont.: McMaster Univ., 1984. (*a*)

———. "Monetary Asset Demand Functions: Theory and Estimates." Mimeographed. Calgary: Univ. Calgary, 1984. (*b*)

Simpson, Thomas D., and Porter, Richard D. "Some Issues Involving the Definition and Interpretation of Monetary Aggregates." In *Controlling Monetary Aggregates III.* Fed. Reserve Bank Boston Conference Ser. no. 23. Boston, Mass.: Fed. Reserve Bank, 1981.

Sims, Christopher A. "Money, Income, and Causality." *A.E.R.* 62 (September 1972): 540–52.

Spindt, Paul A. "Modeling the Monetary Multiplier and the Controllability of the Divisia Monetary Quantity Aggregates." *Rev. Econ. and Statis.* 66 (May 1984): 314–19.

Swamy, Paravastu A. V. B., and Tinsley, Peter A. "Linear Prediction and

Estimation Methods for Regression Models with Stationary Stochastic Coefficients." *J. Econometrics* 12 (February 1980): 103–42

Theil, Henri. *Economics and Information Theory* Amsterdam: North-Holland, 1967.

Tinsley, Peter A., and Garrett, Bonnie, with Friar, Monica E. "The Measurement of Money Demand." Special Studies Paper no. 133. Washington Bd Governors Fed. Reserve System, 1980.

Tinsley, Peter A.; Spindt, Paul A.; and Friar, Monica E "Indicator and Filter Attributes of Monetary Aggregates: A Nit-picking Case for Disaggregation." *J Econometrics* 14 (September 1980): 61–91.

Tobin, James. "The Interest-Elasticity of Transactions Demand for Cash." *Rev. Econ. and Statis.* 38 (August 1956): 241–47

Tornqvist, L. "The Bank of Finland's Consumption Price Index " *Bank Finland Bull*, no. 10 (1936), pp. 1–8

Wall, Kent D. "Estimation of Rational Distributed Lag Structural Form Models." Mimeographed. Cambridge, Mass N B E R , 1974

Job Matching and Occupational Choice

Robert A. Miller

Carnegie-Mellon University

This paper presents a model of job matching that generalizes the existing literature by allowing for different job types, or occupations. Such differences affect the value of job-specific experience, inducing a career profile where certain types are sampled before others. More specifically, the analysis shows that it is optimal for the young and inexperienced to gravitate toward jobs exhibiting a certain kind of risk. Then, after deriving the equilibrium job turnover rate for an economy in which people do not switch occupations, panel data are used to estimate its underlying parameters. The hypothesis that people do not switch occupations is rejected against the alternative that they do, thus providing empirical support for the theoretical extension undertaken here.

I. Introduction

The relationship linking skills that are job specific to turnover decisions has long been considered an important issue in the economics of labor mobility. Earlier writers, most notably Becker (1962, 1975), Mincer (1962), and Oi (1962), perceived a two-way flow between job-specific training and tenure. On the one hand, Becker asserts the "willingness of workers or firms to pay for specific training should . . . closely depend [negatively] on the likelihood of labor turnover" (1962, p. 19). On the other hand, these authors also recognized that the more job-specific capital a worker has, the more destructive a

---

This paper was drawn from the first chapter of my doctoral dissertation (1982). I am most indebted to William Brock and James Heckman, who jointly chaired my committee, Paul Silver, who programmed the computations and provided many useful comments; and two anonymous referees, whose advice led to significant revisions of a previous draft. The research was supported by NSF grant SES-8107963.

[*Journal of Political Economy*, 1984, vol. 92, no. 6]

© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9206-0009$01.50

separation would be, and hence the less likely he is to break the match.

Combined, these two effects yield a decreasing hazard: in other words, the conditional probability of quitting or being laid off falls with greater tenure. For the longer a person has held a job, probably the more specific capital he has acquired (since a worker in whom little capital was invested is more likely to have been filtered out previously). And the higher is his expected output due to job-specific factors, the smaller the chance of a separation occurring.

The decreasing hazard hypothesis is consistent with the observation that certain identifiable groups in the work force turn over more often, experience lower rates of wage growth, and apparently accumulate less specific human capital than others. Furthermore, studying job tenure, several investigators, including Borjas and Rosen (1979), Bartel and Borjas (1981), and Mincer and Jovanovic (1981), have found persistent evidence of positive duration dependence even after they control for socioeconomic characteristics described in the data. That is, among cohorts whom the data measure as being equivalent, the probability of sticking to a job, conditional on past experience, increases with tenure.

Models of job matching and turnover provide a simple framework for analyzing issues related to the accumulation of job-specific capital. An article by Jovanovic (1979) demonstrated that if a person can learn only through specific experience how suited he is to a job, then an (eventually) decreasing hazard emerges from the optimal decision rule. Intuitively, the more experience a person has in a job, the better
the assessment of his compete
affect this assessment and pr

The contributions of Jovan
how optimizing behavior und
generate duration dependen
served heterogeneity (defined
matches known to the individu
data and accordingly unobserved
theless, as Heckman and Borjas (1980) have remarked
context, data on turnover alone cannot help one to decide whether unobserved heterogeneity or the behavioral rules generated by economic models are causing positive duration dependence found by econometricians. Unless assumptions, typically strong and unjustifiable, specify the probability distribution describing unobserved heterogeneity across the population, a statistical test that chooses between the two alternatives is impossible to devise[1]

---

[1] While the discussion of Heckman and Borjas (1980) was concerned with transitions between employment and unemployment, it is clear their conclusions apply more gen-

Thus, work over the last 2 decades has established an interdependence between job-specific capital and labor mobility, formally explained precisely how this interdependence might arise, and produced empirical results that would be broadly consistent with the formal theory but for the presence of unobserved heterogeneity.

The purpose of this paper is to explore the theoretical and empirical ramifications of generalizing the job-matching paradigm. In job-matching models, a Bayesian decision maker combines prior information about each job's characteristics with sample information he has accumulated from working experience on that job in order to forecast its future returns (both pecuniary and nonpecuniary). The next section lays out a framework to accommodate situations where his prior beliefs may differ across jobs. This extension suggests the following definition of an occupation: loosely speaking, two jobs belong to the same occupation if a person's prior beliefs about them are identical. Hence differences between such jobs emerge only through specific experience.

Drawing from the statistical theory of multi-armed-bandit processes, Section III investigates the optimal order to sample jobs and the optimal duration time for each job. (In previous work, the order in which jobs are sampled is immaterial because people have the same prior beliefs over all jobs.) These investigations yield several predictions about life-cycle career patterns. They provide an explanation based on rational behavior for the longstanding empirical regularity, well captured by the expression "young and foolish," that the inexperienced gravitate toward risky ventures. Because younger workers are less experienced than their older counterparts, they more willingly try out jobs where success is rare; older people who have quit them already know they themselves are unsuitable.

A numerical analysis is conducted in Section IV to compute job turnover generated by optimal behavior. The results show that if the quality of job matches is difficult to forecast without specific experience or can be determined quite accurately with only a small amount of experience, turnover is higher.

The empirical parts of this paper, Sections V–VII, focus mainly on two issues. One expects a person's socioeconomic background,

---

erally. The claim is starkly illustrated by considering the sequence of probabilities for breaking a job match at specified tenure levels, derived from sequentially maximizing an *intertemporal utility function subject to a production and information technology* Call this sequence $\{p_i\}_{i=1}^{\infty}$. Contrast this setup with a very simple model in which a population of identical agents work in various jobs $\{i\}_{i=0}^{\infty}$, which are labeled according to the fixed time they take to complete. Suppose the proportion of jobs that last i periods is p_i. Then clearly turnover behavior generated by one model is indistinguishable from the other.

through its effects on his prior beliefs, to determine in part how frequently he switches jobs. Thus, differences observed across socioeconomic groups in layoffs and quits might be attributable to optimal decisions induced by different levels of schooling acting on prior beliefs. Accordingly the first set of hypotheses tests whether schooling is a significant parameter in determining them. The second issue is whether, from an econometrician's viewpoint, each person spends his entire working career in what is defined here as a single occupation. If so, he has the same prior beliefs about the job-matching characteristics of successive employment positions held, and the theoretical portion of this paper would receive scant empirical support. Similarly the practice of ignoring occupational effects, widespread in related work on labor force participation and search unemployment, would be indirectly upheld. A second set of hypotheses addresses this issue.

Section V provides some informal diagnostics on these matters from the Coleman-Rossi data set. Then in Section VI a maximum likelihood estimator is constructed to mesh the turnover data with the probabilities generated numerically for an economy where nobody switches occupations. Maximum likelihood estimates are obtained and hypothesis testing is conducted in Section VII. Overall, the empirical findings support the view that prior beliefs differ across people and also across the jobs any one person works in.

II. The Individual's Problem

Perhaps the easiest way to pic
Crusoe shipwrecked at date
rescue foreseen, he maximiz
sequentially choosing one of
produce it. Through experi
techniques match the geograp
helps him make wiser decisions. It is
will be experimented with first, whether all of them
long it takes Crusoe to settle down with one (if ever), and so on. The transparent interpretation follows from equating food output with a person's income or utility stream, farming techniques with jobs, and the suitability of those techniques to the island's geography with productivity differences of an individual across different jobs.

To help fix ideas, some notation is introduced at this point. Denote by x_{mt} the agent's return, which includes both pecuniary and nonpecuniary aspects, from working in the mth job at time $t \in T \equiv \{0, 1, 2, \ldots\}$, chosen from the job set M. Suppose a return is not directly observed unless the person actually works in the associated job, and furthermore it is the sum of three components:

$$x_{mt} = \psi_t + \xi_m + \sigma_m \epsilon_{mt}. \tag{1}$$

The first term on the right-hand side of (1) denotes a time-trended variable independent of the job and is observed regardless of where the individual works. It may be due to business cycle fluctuations, the effects of age, or general experience and formal education. The second term is a match parameter that is time invariant. The individual does not observe ξ_m directly, but before acquiring any experience in the job he believes it to be normally distributed $N(\gamma_m, \delta_m^2)$. If he had rational expectations, this distribution would describe the latent distribution of talents in that job across the subpopulation of people with socioeconomic characteristics similar to his. The third term is never observed, but σ_m is known and ϵ_{mt} is an independent random variable drawn from the standard normal distribution $N(0, 1)$. The idea behind the second two terms on the right-hand side of (1) is that match quality cannot be ascertained exactly but must be guessed at by combining prior knowledge about matches of a similar type with the returns history from the current match.

Suppose τ_{mt} measures the experience the worker has accumulated in the mth job by period t. In other words, letting d_{ms} be unity if the mth job is chosen at time s and zero otherwise,

$$\tau_{mt} \equiv \sum_{s=0}^{t-1} d_{ms}.$$

Using this notation, the total amount of time spent in the mth job is $\tau_{m\infty}$, which is called completed tenure and henceforth abbreviated to τ_m.

Both the theoretical and empirical portions of this analysis categorize jobs according to whether outsiders can perceive any differences or not. Individuals evaluate jobs in the same category solely on match-specific grounds, whereas when assessing jobs belonging to different categories, both specific and general attributes are relevant. In this sense the model admits a precise definition of an occupation.

DEFINITION 1: Two jobs m and m' belong to the same occupation n if and only if $(\gamma_m, \delta_m, \sigma_m) = (\gamma_{m'}, \delta_{m'}, \sigma_{m'})$. For future reference define their common information factor $\alpha_m \equiv \sigma_m^2 \delta_m^{-2}$. Also denote the set of jobs within the nth occupation by (n).

The definition implies that a person holds the same prior beliefs about the return stream he would receive if he worked in jobs belonging to the same occupation. Furthermore, the rate at which he learns about such matches is identical.

Given the model's specification, beliefs about the mth match can be characterized as a normal distribution $N(\gamma_{mt}, \delta^2_{mt})$ where, following Zellner (1971), a Bayesian updating rule implies

$$\begin{aligned}\gamma_{mt} &= \left[\delta_m^{-2}\gamma_m + \sigma_m^{-2}\sum_{s=0}^{t-1}(x_{ms} - \psi_s)d_{ms}\right](\delta_m^{-2} + \tau_{mt}\sigma_m^{-2})^{-1},\\ \delta_{mt} &= (\delta_m^{-2} + \tau_{mt}\sigma_m^{-2})^{-1/2}.\end{aligned} \tag{2}$$

Observe from equations (2) that beliefs about a match change only with experience on the job.

The agent's objective is to maximize expected returns from jobs over his lifetime by choosing a decision rule d, which maps available information into his preferred job choice at each point in time:

$$d(\{\gamma_{mt}, \delta_{mt}\}_{m\in M}) = \mathop{\times}_{m\in M} d_{mt}(\{\gamma_{mt}, \delta_{mt}\}_{m\in M}),$$

where for all $(m, t) \in M \times T$, $d_{mt} \in \{0, 1\}$ and $\Sigma_{m\in M} d_{mt} = 1$. To summarize, the model is formally characterized by its valuation function,

$$V(\{\gamma_m, \delta_m\}_{m\in M}) = \max_d E_0\left\{\sum_{t=0}^{T}\sum_{m\in M}\beta^t d_{mt} x_{mt}\right\}, \tag{3}$$

where E_t stands for expectations conditional on information $\{\gamma_{mt}, \delta_{mt}\}_{m\in M}$ available at time $t \in T$, and β is the discount factor.

III. Optimal Decision Making

The characterization of the sol
damental (but until recently
(1974).[2] Their theorem on m
this application that each job is
beliefs about its future return
optimal decision is to choose
timality of this simple rule is
picked depends on the characteristic
20 parameters must be kept track of (both the
for match distributions being updated). The dynamic allocation index (DAI) strategy considerably simplifies the computation of the solution because each index depends on the two characteristics of the job associated with it alone.

The definition below precisely specifies the index for this particular application.

Definition 2: Denote by $\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})$ the dynamic allocation index for the mth job given beliefs $(\gamma_{mt}, \delta_{mt})$. It is the real-valued mapping

---

[2] See also Gittins (1979) for a review of work in the area.

$$\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt}) = \sup_{\tau \geq t} \left\{ \frac{E_t\left[\sum_{r=t}^{\tau} \beta^{r-t}(x_{mr} - \psi_r)\right]}{E_t\left(\sum_{r=t}^{\tau} \beta^{r-t}\right)} \right\}.$$

In other words, the index is the valuation function for the following (fictitious) problem: at time t, after netting out the general component ψ_t, choose a stopping time τ to maximize the expected return stream from remaining on the mth job $E_t[\Sigma_{r=t}^{\tau-1} \beta^{r-t}(x_{mr} - \psi_r)]$ divided by the current value of receiving a unitary return over that phase $E_t(\Sigma_{r=t}^{\tau-1} \beta^{r-t})$.

Proposition 1 formally states how the indices are used in computing the optimal decision rule d^o. (All propositions are proved in App. A.)

PROPOSITION 1: Denoting d^o as the optimal decision rule for the problem described by equation (3), if $d_{kt}^o = 1$, then $\mathrm{DAI}_k(\gamma_{kt}, \delta_{kt}) = \max_{m \in M}[\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})]$. Furthermore, any tie-breaking rule between the maxima is optimal.

The properties of the solution are introduced by briefly characterizing a subcase that appears in the literature on optimal search. Setting $\sigma_m = 0$ for all $m \in M$ and $\psi_t = 0$ for all $t \in T$ specializes this framework to the normal variant of Weitzman's (1979) analysis of Pandora's problem. (If in addition $(\gamma_m, \delta_m) = (\gamma_{m'}, \delta_{m'})$ for all $\{m, m'\} \subseteq M$, a prototype search model of variety reviewed in Lippman and McCall [1976] is generated.) Equation (1) becomes

$$x_{mt} = \xi_m. \tag{4}$$

Definition 1 now implies that the index value of a sampled job is just its (constant) return. From proposition 1 a very simple career strategy emerges. At date zero rank the indices from highest to lowest and sample jobs, one per period, using this ordering, until the maximum return received in any previous period exceeds the highest-ranking index value among the remaining unsampled jobs; then work in the job associated with this maximal return forever.

Deriving the index values for this specialization is relatively straightforward. Consider a (perhaps hypothetical) job whose match is drawn from a standard normal distribution $N(0, 1)$. Call its index value $D(0, \beta)$. (The significance of the zero will presently become apparent.)

DEFINITION 3: For $\sigma_m = 0$ define $D(0, \beta) \equiv \mathrm{DAI}_m(0, 1)$.

For this specialization all job indices are a linear combination of

$D(0, \beta)$. Each index is additive in the expected return γ_m and another component $\delta_m D(0, \beta)$, the latter increasing proportionately with δ_m.

PROPOSITION 2: If $\sigma_m = 0$ then $\mathrm{DAI}_m(\gamma_m, \delta_m) = \gamma_m + \delta_m D(0, \beta)$.

The characterization is completed by utilizing proposition 3 to compute $D(0, \beta)$ numerically. (The top line of table B1 in App. B depicts the mapping $D(0, \beta)$, which is monotone increasing in β.)

PROPOSITION 3: Denote by $\Phi(\epsilon)$ the standard normal cumulative distribution function and $\Phi'(\epsilon)$ the associated density. Then for each $\beta \in (0, 1)$ the real number $D(0, \beta)$ uniquely solves $D(0, \beta) = \{1 - \beta\Phi[D(0, \beta)]\}^{-1}\beta\Phi'[D(0, \beta)]$.

The two preceding paragraphs fully operationalize the index rule for a particular specialization. The remainder of this section analyzes its general properties and discusses the economic implications. An implication of proposition 1 is that a person choosing job $m \in (n)$ will not later switch to another occupation $n'' \neq n$ at least until all jobs $m' \in (n)$ have been sampled. The reasoning is as follows: When the worker initially picks $m \in (n)$, if at that time the person has no experience in $m' \in (n)$ either, from definition 1 the index values for m and m' are identical. Also, because the worker prefers m to every job $m'' \in (n'')$, by proposition 1 the index value for m'' is less than the index value for m, and thence m'. But from equations (2) and definition 2 index values cannot change without specific experience. Therefore, appealing to proposition 1 again, m'' will not optimally be chosen unless the index value for m' falls as a result of work experience (in m').

COROLLARY 1: Let $m \in (n)$ and $m'' \in (n'')$ where $n \neq n''$. If $d''_{mt} = d''_{ms} = 1$ for some $s > t$, then for all $m' \in (n)$ [illegible] 0.

Suppose, after acquiring ex
is sampled. The propositio
return to m until every job
picked, from definition 1, th
$\in (n)$ if he has no experienc
common index value for m
Using the same argument ap
experience in m'' must be acquired
bly fall below the index value for m.

COROLLARY 2: Suppose $\{m, m', m''\} \subseteq (n)$. If $d_{mt} = d_{ms} = d_{mr} = 1$ for some $t < s < r$, then $\tau_{m''r} > 0$.

Corollary 1 says that whenever a new occupation is entered, all jobs within it are sampled before any occupational shift occurs. From corollary 2 multiple spells within a job will not occur at least until some experience has been acquired in all jobs belonging to the same occupation. So if there are an infinite number of jobs in the first occupation a worker enters, all jobs he ever takes last one spell and belong to that occupation. For this specialization the model essentially reduces to Jovanovic's (1979) framework; heterogeneity across the population becomes the sole reason for people to work in different occupations.

COROLLARY 3: Let $n(\#)$ denote the number of jobs belonging to occupation n. If $m' \in (n)$, $n(\#) = \infty$, and $d_{m'0} = 1$, then

$$\tau_m = \begin{cases} \min\limits_{s \in T}\{\tau_{ms} | d_{ms} = 1, d_{m(s+1)} = 0\} & m \in (n) \\ 0 & m \notin (n). \end{cases}$$

The properties of the optimal decision rule d^o can be further determined by examining the properties of the mappings $\mathrm{DAI}_m(\gamma_m, \delta_m)$. In the present application, an index can be expressed more simply than in definition 2. Central to this simplification is the standard index $D(\sigma_m^2, \beta)$.

DEFINITION 4: For all (σ_m, β) define $D(\sigma_m^2, \beta) \equiv \mathrm{DAI}_m(0, 1)$.

Less formally, the standard index is the DAI for a job whose match parameter ξ_m is drawn from the standard normal distribution and whose return net of the general component $(x_{mt} - \psi_t)$ is perturbed by a normal random variate ϵ_{mt} with variance σ_m^2. Whether such a job actually exists is immaterial; the purpose of inventing the notation is to provide a reference point against which all indices are judged.

PROPOSITION 4: $\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt}) = \gamma_{mt} + \delta_{mt} D(\alpha_m + \tau_{mt}, \beta)$.

Intuitively, the index for any job $m \in M$ is just the sum of its expected return (net of the general component) plus a term $\delta_{mt} D(\alpha_m + \tau_{mt}, \beta)$, which represents the value of the job as a source of information about itself. (Recall from definition 1 that α_m is the information factor.) Clearly the properties of $\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})$ follow from those of $D(\alpha_m + \tau_{mt}, \beta)$, since given $(\gamma_{mt}, \delta_{mt})$, the former is just a linear transformation of the latter, and although $D(\alpha_m + \tau_{mt}, \beta)$ does not admit a closed-form representation, it can be computed numerically.

Appendix B describes the algorithm for computing $D(\alpha_m, \beta)$. It is based on the following proposition.

PROPOSITION 5: Let F denote the space of bounded continuous real valued functions taking their domain on the space $(-\infty, \infty) \times (0, \infty)$. For all $f \in F$ define C, a mapping from F to itself, as

$$C[f(\gamma, \alpha)] = \beta \int_{-\infty}^{\infty} \max\Big(0, (1 - \beta)\{\gamma + \epsilon[\alpha(\alpha + 1)]^{-1/2}\}$$

$$+ f\{\gamma + \epsilon[\alpha(\alpha + 1)]^{-1/2}, \alpha + 1\}\Big) d\Phi(\epsilon).$$

There exists a unique function $g(\gamma, \alpha) \in F$ satisfying $g(\gamma, \alpha) = C[g(\gamma, \alpha)]$. Moreover, defining $C_k(f)$ recursively as $C[C_{k-1}(f)]$, for all $f \in F$, $\|C_k(f) - g\| \leq (1 - \beta)^{-1}\|C_k(f) - C_{k-1}(f)\|$. Finally for all (α, β)

$\in (0, \infty) \times (0, 1)$ the standard index solves $D(\alpha, \beta) = (1 - \beta)^{-1}\alpha^{1/2}g[-\alpha^{-1/2}D(\alpha, \beta), \alpha]$.

Briefly, the function $g(\gamma, \alpha)$ is approximated by recursively applying C to the zero function. Since the final inequality uniformly bounds the approximation error, this method provides the means for computing g to an arbitrary degree of accuracy. The standard index $D(\alpha, \beta)$ is then solved for each (α, β) from the equation in g, using a fixed-point routine.

Figure 1 illustrates the qualitative features obtained from a numerical analysis of $D(\alpha_m, \beta)$ described in Appendix B (which also reports the results in table B1). In the top right quadrant, the standard index $D(\alpha_m, \beta)$ is mapped as a function α_m for a particular value of β. On the left side of the horizontal axis, the product $\delta_{mt}D(\alpha_m + \tau_{mt}, \beta)$ is graphed, while the lower portion of the vertical axis is, from proposition 4, $\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})$.

From an economist's perspective, six main results emerge. The current job yields two benefits: payoff from expected reward for the period and information about future returns within that particular job. Naturally, a person will opt for the job with the bigger expected reward if the value from the informational component is the same for the two alternatives. This first result is reflected in the third quadrant of figure 1. Raising γ_{mt} lowers the intercept on the q_1 axis.

Supposing expected current returns are equal, he would choose the job that provides the most valuable information. Information value is embodied within the term $\delta_{mt}D(\alpha_m + \tau_{mt}, \beta)$ It depends on three factors: what is already known, given by the precision δ_{mt}^{-2}, the rate of acquisition measured by σ_m, and pr[illegible]
one's initial information abo[illegible]
less one is prepared to forg[illegible]
Its attraction as a source [illegible]
tighter priors. This is the s[illegible]
would reduce both the slop[illegible]
also α_m (from def. 1), the lo[illegible]
Hence the product $\delta_m D(\alpha_m, \beta)$ [illegible]

Furthermore, greater variability [illegible]
tards learning about the underlying process, that is, prior beliefs are not affected as much by any given amount of on-the-job experience when the latter is an unreliable guide for measuring competence in an occupation. That the informational component is therefore less valuable constitutes a third result Diagrammatically, if σ_m increases, α_m increases and $D(\alpha_m, \beta)$ falls.

Fourth, specific experience τ_{mt} reduces $D(\alpha_m + \tau_{mt}, \beta)$ at a decreasing rate. Also, from equation (2), δ_{mt} is decreasing in τ_{mt}. Accordingly, the ray from the origin in the top left quadrant swivels clockwise as τ_{mt}

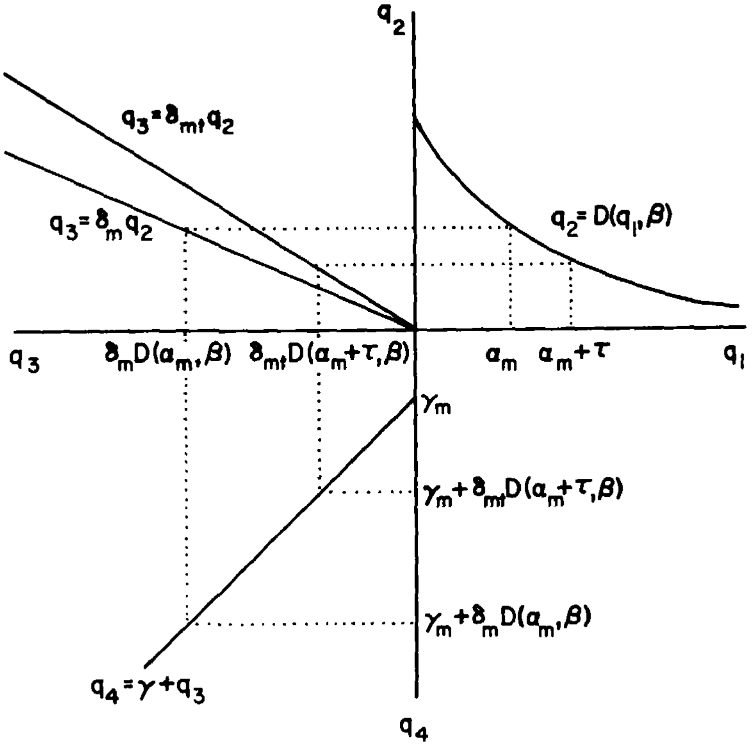

q_2
$q_3 = \delta_{mt} q_2$
$q_3 = \delta_m q_2$
$q_2 = D(q_1, \beta)$
q_3
$\delta_m D(\alpha_m, \beta)$
$\delta_{mt} D(\alpha_m + \tau, \beta)$
α_m
$\alpha_m + \tau$
q_1
γ_m
$\gamma_m + \delta_{mt} D(\alpha_m + \tau, \beta)$
$\gamma_m + \delta_m D(\alpha_m, \beta)$
$q_4 = \gamma + q_3$
q_4

FIG. 1.—The dynamic allocation index

increases. Consequently the product $\delta_{mt} D(\alpha_m + \tau_{mt}, \beta)$ falls. Thus the value of additional information declines as it is acquired.

In addition, matches about which a precise estimate can be ascertained more quickly lose their attractiveness as an information source sooner. To see this point consider the limiting case when a single-period tryout perfectly reveals match quality, that is, when $\sigma_m = 0$. With reference to the top left quadrant, as σ_m approaches zero, $(\delta_m^{-2} + \sigma_m^{-2})^{1/2}$ approaches infinity, so using (2) the ray from the origin $q_2 = \delta_{m1} q_3$ becomes vertical, folding into the q_2 axis and making the index value $\mathrm{DAI}_{m1}(\gamma_{m1}, 0)$ equal to the true match parameter ξ_m. Obviously, this would not occur if $\sigma_m > 0$.

Finally, people who discount future payoffs more value information less. Intuitively, this result follows directly from the fact that information about a job match is a form of human capital. In the diagram, decreasing β reduces the standard index $D(\alpha_m, \beta)$ for all α_m.

These findings yield predictions about life-cycle behavior and the composition of labor markets. A theory of equalizing differences would imply that jobs with high informational benefits pay less on average in equilibrium. They attract the inexperienced who relatively quickly discover their personal match. Workers who have passed through the experimentation stage will not be found in such jobs unless their match is uncommonly good. Jobs yielding a great deal of information about matches therefore tend to be filled mainly by a transient inexperienced population, a cross-sectional average of whom earn modest returns, plus a much smaller number of experienced, very successful, permanent workers. The performing arts sector probably fits into this category.

On the other hand, jobs with fewer informational benefits tend to exhibit less turnover, higher-than-average returns to entrants, and a greater concentration of incomes about the mean.

A young worker is therefore more likely to quit his job for two reasons. First, there is a lower probability that within a particular occupation his match will be superior. Typically an older worker has had more time to find one. Second, a greater percentage of the young are found in occupations where high turnover is endemic. Because they are less self-aware than older people, young people are more willing to experiment in activities that increase their self-awareness.

Ever since Adam Smith observed the attraction to youth of some occupations yielding risky returns, he and others following him have explained the phenomenon in terms of a taste for the uncertain, overoptimistic expectations about ability, or status.[3] This section has demonstrated what types of uncertainty attract the inexperienced, explaining their attraction in terms [illegible] without recourse to tastes or ir[illegible]

IV. Job Turnover

Roughly speaking, job turno[illegible] and match-specific factors [illegible] ranking of index values, the p[illegible] some future time potentially depen[illegible] jobs. Therefore, a complete characterization of equilibrium turnover rates presents one with a formidable task, requiring the numerical computation of sequences of multiple integrations over the belief space $\{\gamma_{mt}, \delta_{mt}\}_{m \in M}$. Nevertheless some progress can be made at moderate cost.

---

[3] "The contempt of risk and the presumptuous hope of success, are in no period of life more active than at the age at which young people chuse their professions" (Smith 1937, p. 109). In a similar vein, Marshall (1948) writes, "If an occupation offers a few extremely high prices, its attractiveness is increased out of all proportion to their aggregate value. For this there are two reasons. The first is that young men of an adventurous disposition are more attracted by the prospects of a great success than they are deterred by the fear of failure, and the second is that the social rank of an occupation depends more on the highest dignity and the best position which can be attained through it than on the average good fortune of those engaged in it" (1948, p. 554).

Considering job $m \in (n)$ held at time $t \in T$, suppose there exists another one belonging to the same occupation $m' \in (n)$ in which the person has no experience. That is, $\tau_{m't} = 0$. By corollary 2 the mth match has not previously been interrupted: at time t the worker is still in his first spell. By corollary 1 if and when he quits, he will switch to a new job in (the same) occupation n. This section analyzes the discrete hazard function associated with such transfers. They are especially significant when $n(\#) = \infty$. For appealing to corollary 3, the mth match lasts only one spell, and hence in that case the discrete hazard for the first spell is identically the hazard for the job match.

DEFINITION 6: $h_{m\tau} \equiv P\{\gamma_{mt} + \delta_{mt}D(\alpha_m + \tau_{mt}, \beta) < \gamma_m + \delta_m D(\alpha_m, \beta) | \tau_{mt} = \tau\}$.

Let $m \in (n)$. The statement $\tau_{mt} = \tau$ implies from proposition 1 that, if $\tau_{m't} = 0$ for some $m' \in (n)$, given optimization, the inequality $\gamma_{ms} + \delta_{ms}D(\alpha_m + \tau_{ms}, \beta) \geqslant \gamma_m + \delta_m D(\alpha_m, \beta)$ holds for all $\tau_{ms} < \tau_{mt}$. The fact that at time t this inequality is reversed implies that it is no longer optimal to choose m over m'.

When $\sigma_m = 0$, match quality is perfectly revealed after one period's experience; hence $\mathrm{DAI}_m(\gamma_{m1}, \delta_{m1}) = \mathrm{DAI}_m(\gamma_{m1}, 0) = \xi_m$. Noting that beliefs above ξ_m are initially normally distributed $N(\gamma_m, \delta_m^2)$, proposition 2 and definition 6 imply, for this specialization,

$$h_{m1} = P\{\xi_m < \gamma_m + \delta_m D(0, \beta)\} = \Phi[D(0, \beta)].$$

Furthermore, since no information is acquired after working more than one period, $h_{m\tau} = 0$ for all $\tau > 1$.

To derive the discrete hazard for cases where match quality is only imperfectly revealed by returns, the probability distribution $\Psi_{mt}(\rho)$ of the transformed means $(\gamma_{mt} - \gamma_m)\sigma_m^{-1}$ is required.

DEFINITION 7: $\Psi_{m\tau}(\rho) \equiv P\{\gamma_{mt} < \gamma_m + \rho\sigma_m | \tau_{mt} = \tau\}$.

The hazard $h_{m\tau}$ and the distribution $\Psi_{m\tau}(\rho)$ may be recursively computed for successive values of $\tau \in T$. First the distribution $\Psi_{m1}(\rho)$ is found by updating the prior transformed mean (of zero) with a realization from the standard normal distribution $\Phi(\epsilon)$. If the transformed posterior mean lies below the critical value $[\alpha_m^{1/2}D(\alpha_m, \beta) - (\alpha_m + 1)^{-1/2}D(\alpha_m + 1, \beta)]$ the spell ends. Hence the proportion of spells ending after one period's experience h_{m1} is $\Psi_{m1}[\alpha_m^{1/2}D(\alpha_m, \beta) - (\alpha_m + 1)^{-1/2}D(\alpha_m + 1, \beta)]$. The probability distribution $(1 - h_{m1})^{-1}[\Psi_{m1}(\rho) - h_{m1}]$ now forms the distribution of prior means for beliefs about remaining, uncompleted spells with one period's tenure. The procedure outlined above is repeated to find $\Psi_{m2}(\rho)$, and so on.

Proposition 6: For all $(m, \tau) \in M \times T$, if $\sigma_m > 0$, then

$$h_{m\tau} = \psi_{m\tau}[\alpha_m^{-1/2} D(\alpha_m, \beta) - (\alpha_m + \tau)^{-1/2} D(\alpha_m + \tau, \beta)]$$

$$\psi_{m1}(\rho) = \Phi[\rho\alpha_m^{1/2}(\alpha_m + 1)^{1/2}]$$

$$\psi_{m(\tau+1)}(\rho) = (1 - h_{m\tau})^{-1}\left\{\int_{-\infty}^{\infty} \psi_{m\tau}[\rho - \epsilon(\alpha_m + \tau)^{1/2} \times (\alpha_m + \tau + 1)^{1/2}]d\Phi(\epsilon) - h_{m\tau}\right\}.$$

Appendix C describes how the sequence $\{h_{m\tau}\}_{\tau=1}^{\infty}$ was computed over a grid (α_m, β) lying in the space $[0.1, 17.1] \times [0.1, 0.95]$ from values obtained for the standard index $D(\alpha_m, \beta)$ and reports the results. For any (α_m, β) coordinate pair the sequence declines monotonically and has a long tail for reasons explained in the Introduction. Three further results are depicted in table C1.

First, the more future payoffs are discounted, the lower the hazard, because the value of finding a superior job match is diminished. Thus, lower values of β uniformly shift the hazard down.

Second, increasing the information factor α_m lowers the hazard. Recall from definition 1 that $\alpha_m \equiv \sigma_m^2\delta_m^{-2}$. Therefore if the quality of job matches is difficult to forecast without specific experience (meaning δ_m is high) or can be determined quite accurately from just a small amount of working experience (which occurs when σ_m is low), turnover is higher.

Third, over 50 percent separate with tenure of 1 period.

Corollary 4: $h_{m1} = \Phi\{[D(\alpha_m, \beta)$

The economic intuition b
ward. By definition 1 an agen
belonging to the same occu
information source declines
of information about match
rational expectations), the no
is unfavorable half the time
fall than rise. When this happ
time he chooses the occupation

V. Data

When any theory is confronted with data, questions about how to calibrate its primitives immediately arise. Here, three theoretical concepts have empirical counterparts in the data, namely jobs, occupations, and returns.

The only type of learning admitted by the model is about job matches. Ideally an empirical application should therefore view jobs in terms of those activities where the most task-specific learning takes place. This point is best made by example. Given the discipline musical scores impose on concert renditions, a violinist's most important match is neither to his orchestra nor to his conductor, but to his instrument. Therefore, treating successive orchestras or conductors as separate job matches would yield poorer empirical results than if the position of violinist itself were regarded as one job match. Needless to say, the detail required to make such an approach operational is not typically available, and the crude employment categories utilized in this study severely limit the inferences one can make.

Two interpretations compatible with the model are that the return x_{mt} represents the utility a person gets in period t from working in job m after purchasing consumption goods with his pecuniary income, or a monetary return plus a value of on-the-job amenities implicitly purchased. The first interpretation is valid if financial markets are closed to agents, while the second applies if the supply of credit is perfectly elastic. Much applied work measures returns with wages, but their link to both interpretations is somewhat tenuous. First, the nonpecuniary aspects of a job, an important consideration in a matching context, are ignored. Second, if financial markets are open, as Jovanovic (1979) notes, the wage contract is nonunique, so wages do not necessarily correspond to realized output. For these reasons wage data were not used; this also means the econometric results are compatible with either theoretical interpretation mentioned above.

The Coleman-Rossi data set was used to quantify the theory and assess its empirical significance. A detailed summary of the data, previously used by Bartel (1979) in her work on migration, has been provided by Ornstein (1976). Very briefly, information was gathered in the winter of 1969 through one interview with each of 822 white and 767 nonwhite men. Among other things, these men were questioned extensively about their entire work history, education, and family background. Appendix D describes the criteria met by the 467 white men making up the sample used in this study.

Table 1 illustrates tenure in and turnover between spells by employment group and education level. Here tenure is measured biennially.[1] The number of people belonging to the sample who were working in each group when the survey was taken is recorded in column 1. Columns 2–4 successively measure the number of people at different tenure levels in their spell by employment group and educational attainment as a percentage of the corresponding total in

---

[1] Ideally the period length (which determines the arrival rate of new information) should be simultaneously estimated along with the other parameters of interest, but this extension remains a topic for future research.

TABLE 1

TENURE AND TURNOVER BY EMPLOYMENT AND EDUCATION

	CURRENT POSITION				PAST SPELLS			
		Percentage with Tenure of				Empirical Hazard		
	Number	≥ 2	≥ 3	≥ 4	Number	1	2	3
Employment:								
Professional	67	76	65	31	183	61	49	65
Farm owner	22	95	90	9	44	55	50	30
Manager	80	80	73	33	128	60	55	61
Clerk	40	82	67	35	175	69	55	44
Salesman	27	77	62	29	138	64	51	54
Craftsman	107	81	65	25	379	61	53	59
Operative	84	80	78	39	553	68	59	53
Serviceman	13	92	61	46	60	73	63	33
Farm laborer	6	83	83	33	144	72	54	63
Nonfarm laborer	21	76	57	33	281	78	55	39
Education:								
Grade school	177	84	75	28	779	70	55	64
High school	113	81	67	33	566	68	58	42
College	84	76	67	35	163	61	50	50

column 1. For example, of the 67 professionals interviewed, 76 percent had current tenure of at least 2 biennial periods (4 years) in 1969. Column 5 is the number of previous, or completed, spells held by the sample population, while the last three columns, the empirical hazards, categorize them by spell length. They show the number of quits at specified tenure levels as a perce [illegible] at least as long. Thus 58 percent [illegible] school graduate and lasting at [illegible]

Some indication of the und [illegible] groups is given by these empi [illegible] occupation in an economy po [illegible] sponding percentages would be [illegible] at this highly aggregated level, [illegible]

The apparent differences in table 1 [illegible] across employment categories could result from the existence of several occupations associated with those groupings. Alternatively, such differences might be wholly attributable to a heterogeneous population. Under the latter hypothesis, job turnover rates depend on lifetime socioeconomic characteristics, for which employment groups serve as a proxy. One interesting specialization of this hypothesis is that people with common socioeconomic characteristics belong to the same employment group. If true, no switching between employment groups would occur. A matrix of relative frequencies depicting transi-

TABLE 2

TRANSITIONS WITHIN AND BETWEEN EMPLOYMENT GROUPS

	Professional	Farm Owner	Manager	Clerk	Salesman	Craftsman	Operative	Serviceman	Farm Laborer	Nonfarm Laborer
Professional (183)	67	1	11	4	4	5	5	1	0	1
Farm owner (44)	0	25	2	2	2	9	39	2	14	5
Manager (128)	11	2	39	4	20	10	9	1	1	3
Clerk (175)	10	0	14	33	7	11	15	2	0	7
Salesman (138)	1	1	27	6	30	9	17	4	0	5
Craftsman (379)	5	0	7	6	5	18	18	2	2	7
Operative (553)	4	3	5	6	4	19	38	3	4	14
Serviceman (60)	3	0	5	8	7	10	30	18	3	15
Farm laborer (144)	2	8	1	1	2	8	28	2	31	16
Nonfarm laborer (281)	1	2	2	8	2	18	40	3	1	22

tions between employment groups was accordingly constructed from the job changes indicated by the data. In cell (i, j) of table 2 (where i designates the row and j the column) is the percentage of completed spells belonging to the ith employment category that were followed by a spell in the jth. On dividing the matrix elements by 100, the specialization above yields an identity matrix, which is plainly rejected by table 2.

A third summary measure of the data is the spell count per job. If there were only one occupation and if jobs (as recorded by the data) were synonymous with matches, from corollary 3 multiple spells would never be observed. However, table 3 exhibits the contrary.

Notwithstanding the highly aggregated form of the data, these

TABLE 3

SPELLS NOT INVOLVING NEW JOBS (by Employment Group)

Professional	Farm Owner	Manager	Clerk	Salesman	Craftsman	Operative	Serviceman	Farm Laborer	Nonfarm Laborer
10	12	8	3	4	13	19	1	1	18

diagnostics provide weak, informal evidence against the hypothesis of a one-occupation economy (as given in def. 1). The empirical hazards differ markedly across the various categories, people do switch groups, and multiple spells within jobs occur.

Several caveats are, however, in order. Although half the spells recorded as completed end within a single (2-year) period and the empirical hazard for completed matches declines, the relative frequencies are not directly comparable with data in Appendix C, because the former are computed from censored data. Moreover, the top portion of table 1 makes no allowance for population heterogeneity. For this reason the probability transition matrix has limited value too: it is hardly surprising that the specialization is not confirmed by the data, given its unreasonable premises. Finally, the assumption implicit in table 3 that positions are synonymous with matches may be questioned; returning to a position previously held can entail another boss, more modern machinery, and/or different duties. In such cases it might be reasonable to presume a new match is beginning.

VI. Econometrics

The pitfalls of interpreting da
score the value of proceeding
aim here is twofold. First, th
occupation model are estimated b
merical analyses already undertaken
hazards depicted in table 1, the hazards estimated below control for both observed and unobserved heterogeneity. Moreover, they are not subject to censoring bias. Second, likelihood ratio tests are conducted to determine the significance of heterogeneity and also whether incorporating employment group dummies increases the explanatory power of the model; under the null hypothesis of one occupation the collective contribution of the employment dummies would be insignificant.

The discrete hazards, as given by definition 6, form the basis of the likelihood function L. Let $p_\tau(\alpha, \beta)$ denote the unconditional probability of a person with discount factor β working τ periods in a new job with information factor α before switching to another new job in the same occupation. That is,

$$p_\tau(\alpha, \beta) \equiv h_\tau(\alpha, \beta) \prod_{s=1}^{\tau-1} [1 - h_s(\alpha, \beta)].$$

Then the joint probability of spell duration times observed across the sample is

$$L = \prod_{j \in J} \prod_{i \in (j)} \left\{ \sum_{\tau \in s(i)} [p_\tau(\alpha_{1i}, \beta_{1j})\lambda + p_\tau(\alpha_{2i}, \beta_{2j})(1 - \lambda)] \right\}.$$

There are J workers, and associated with the jth person are (j) job spells; in the one-occupation model each spell $i \in (j)$ corresponds to a different match. Denote by $s(i)$ the information contained in the data set about the duration of the spell associated with the ith observation. For the ith observation, the econometrician might observe a spell ending at tenure τ; then $s(i) = \tau$. So far as the current spell is concerned, the econometrician observes it lasting τ periods or more; in this case $s(i) = \{\tau, \tau + 1, \ldots\}$.

Assuming momentarily that the parameters (α_i, β_j) are either $(\alpha_{1i}, \beta_{1j})$ or $(\alpha_{2i}, \beta_{2j})$, for some $\lambda \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$, the probability of the event $s(i)$ is $\Sigma_{\tau \in s(i)}[p_\tau(\alpha_{1i}, \beta_{1j})\lambda + p_\tau(\alpha_{2i}, \beta_{2j})(1 - \lambda)]$. Conditional on $\{\alpha_{1i}, \alpha_{2i}, \beta_{1j}, \beta_{2j}\}_{j \in J}$, the maximum likelihood estimate for λ can be obtained from L. This approach explicitly recognizes that certain variables relevant in determining match quality, like physical stamina, emotional makeup, and work environment, are measured by the data only poorly, if at all. Their joint effect is modeled here as a Bernoulli distribution. So with probability λ the ith spell falls into the first unobserved state, and with probability $(1 - \lambda)$ into the second.[5]

Naturally, the set $\{\alpha_{1i}, \alpha_{2i}, \beta_{1j}, \beta_{2j}\}_{i \in (j)}$ is not observed for any $i \in (j)$, although vectors of dummy variables z_i and z_j describing the socioeconomic characteristics of each spell and each person are observed. Accordingly, the information and discount factors are assumed to be linear mappings of these observed characteristics:

$$\begin{aligned} \alpha_{ki} &= z_i' a_k, & k &= 1, 2 \\ \beta_{kj} &= z_j' b_k, & k &= 1, 2. \end{aligned} \tag{5}$$

---

[5] This form of unobserved heterogeneity has a mover-stayer interpretation. Alternatively it may be viewed as an approximation to the nonparametric methods proposed by Heckman and Singer (1982, 1984). Given the dimension of the parameter space and the model's structure, a two-point distribution with interactions was found reasonable for the sample size under consideration.

TABLE 4

MAXIMUM LIKELIHOOD ESTIMATES FOR A ONE-OCCUPATION ECONOMY

	STRUCTURAL PARAMETERS				
EDUCATION	α_1	α_2	β_1	β_2	λ
Grade school	3.1	16 1	.95	.95	7
High school	3.1	15 1	95	95	9
College	16.1	16 1	95	.95	

Substituting (5) into L, maximum likelihood estimates for a_1, a_2, b_1, b_2, and λ are found by evaluating the likelihood at all points belonging to the discrete parameter space depicted in table C1.

VII. Empirical Findings

Table 4 presents the maximum likelihood estimates for a one-occupation economy. Job matches are classified according to the educational attainment of the worker involved, and unobserved heterogeneity is modeled as a binary random variable. Two findings are noteworthy.

First, the effect of education is nonmonotone. After margining out the unobserved heterogeneity, the maximum likelihood estimates of the information factor are 7, 4, and 16 for the respective educational groups. This result is not surprising. Higher educational attainment is conceivably associated with faster learning about match quality, greater opportunities for speciali
choices. On the other hand, ge
for low aptitude in decipherin
tively lower σ, raise δ, lower δ, a
α ambiguous.

Second, the estimates of the
furthermore equal to the uppe
misspecification. Because the theory a
due to poor matching, turnover attributable to other causes probably induces upward bias in the estimated hazards (which are determined from table C1 using the estimates of table 4). Since the hazard is increasing in the discount factor, such upward bias might be responsible for the unrealistically high estimates of β obtained.

To gauge the importance of education and the viability of the assumption that people never switch occupations, seven other variations on the basic framework were run. Four permutations of the model are generated by successively omitting and including education and/or unobserved heterogeneity. In addition, the information factor was

TABLE 5

CONTROLLING FOR EDUCATION ($\chi^2_{.05}(q)$, LRTS)

EMPLOYMENT	UNOBSERVED HETEROGENEITY No	UNOBSERVED HETEROGENEITY Yes
No	(9.5, 792.4)	(18.3, 785.9)
Yes	(33.9, 796.8)	(64, 809.0)

TABLE 6

CONTROLLING FOR EMPLOYMENT GROUP ($\chi^2_{.05}(q)$, LRTS)

EDUCATION	UNOBSERVED HETEROGENEITY No	UNOBSERVED HETEROGENEITY Yes
No	(16.9, 77.3)	(28.9, 69.3)
Yes	(40.1, 81.7)	(66, 81.5)

allowed to depend on the employment group in four of the eight variations.

Table 5 depicts the results of likelihood ratio tests that examine the significance of education. The data support the view that educational attainment is associated with different beliefs and learning speeds, generating systematic differences in turnover behavior. The second entry in each cell LRTS (likelihood ratio test statistic) is minus twice the log likelihood ratio, while the first is the corresponding boundary for a .05 size test calculated from a χ^2 distribution adjusted for the appropriate degrees of freedom, q. Regardless of whether unobserved heterogeneity is accounted for or whether employment dummies are included, the null hypothesis of no effect is (asymptotically) rejected.[6]

Perhaps the main finding in this section is that employment group dummies have an impact. Table 6 shows that, regardless of whether unobserved heterogeneity and educational differences are incorporated, the null hypothesis—that the information factor does not depend on the employment category of the match—is rejected. Since individuals are mobile between employment categories (as seen from table 2), explaining this dependence in terms of unobserved heterogeneity, while a logical possibility, seems implausible. More reasonable is the conclusion that the employment categories, rough as they are, actually indicate something about the nature of the work involved and that the characteristics of matches (or, more generally, job-specific capital) differ across these categories. Given this conclusion, it follows that rational individuals will consider the sorts of issues analyzed by the theory above as they move through their working lives.

---

[6] It may also be worth mentioning that when unobserved heterogeneity is introduced in the presence of schooling and employment group dummies the improvement in the maximum value of the log likelihood is negligible (from −2,202 to −2,201)

VIII. Conclusion

This paper extends the matching literature by formally incorporating the notion that people try out several jobs, perhaps from different occupations, depending on how things go. It exposits the optimal decision rule, showing which jobs should be sampled first and for how long. Then the turnover rate is derived numerically for an economy where, after controlling for socioeconomic traits, all differences between jobs are attributable to match-specific considerations. The parameters of such an economy are estimated using panel data, and the hypothesis that people do not switch occupations is rejected. The empirical findings bolster the theoretical portion of the paper. It argues that since the value of specific experience varies across job types, optimizing behavior induces a stochastic career profile. jobs yielding returns that are subject to a particular form of uncertainty are experimented with first. To the casual social commentator, these jobs seem underpaid, and the young, inexperienced workers engaged in them look gullible or foolhardy; but in point [illegible] from the truth.

Appendix A

This appendix proves the propos[illegible]
frequent use is made of the follo[illegible]

Lemma 1: For all $m \in M$

$$\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt}) = \sup_{\tau > t} E_t\Big[(1 - \beta)\sum_{s=t}^{\tau} \beta^{s-t}(x \ldots$$

Proof: From definition 2,

$$\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt}) = \sup_{\tau > t}\left\{\frac{E_t\left[\sum_{s=t}^{\tau} \beta^{s-t}(x_{ms} - \psi_s)\right]}{E_t\left(\sum_{s=t}^{\tau} \beta^{s-t}\right)}\right\}$$

$$0 = \sup_{\tau \geq t} \left\{ \frac{E_t\left[\sum_{r=t}^{\tau} \beta^{r-t}(x_{mr} - \psi_r)\right]}{E_t\left(\sum_{r=t}^{\tau} \beta^{r-t}\right)} - \mathrm{DAI}_m(\gamma_{mt}, \delta_{mt}) \right\}$$

(subtracting $\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})$ from both sides)

$$= \sup_{\tau \geq t} E_t\left\{\left[\sum_{r=t}^{\tau} \beta^{r-t}(x_{mr} - \psi_r)\right] - \left(\sum_{r=t}^{\tau} \beta^{r-t}\right)\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})\right\}$$

(multiplying through by $[E_t \, \Sigma_{r=t}^{\tau} \, \beta^{r-t}]$)

$$= \sup_{\tau \geq t} E_t\left\{\left[\sum_{r=t}^{\tau} \beta^{r-t}(x_{mr} - \psi_r)\right] - (1 - \beta)^{-1}(1 - \beta^{\tau - t + 1})\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})\right\}$$

since $(1 - \beta)^{-1}(1 - \beta^{\tau - t + 1}) = \Sigma_{r=t}^{\tau} \, \beta^{r-t}$. The proof is completed upon multiplying through by $(1 - \beta)$ and adding $\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})$ to both sides.

PROPOSITION 1: Denoting d^0 as the optimal decision rule for the problem described by equation (3), if $d^0_{kt} = 1$, then

$$\mathrm{DAI}_k(\gamma_{kt}, \delta_{kt}) = \max_{m \in M}[\mathrm{DAI}(\gamma_{mt}, \delta_{mt})].$$

Furthermore, any tie-breaking rule between the maxima is optimal.

PROOF OF PROPOSITION 1. Appealing to lemma 1 and noting that the general component ψ_t is received regardless of which job $m \in M$ is chosen, theorem 2 of Gittins and Jones (1974) applies directly.

PROPOSITION 2: If $\sigma_m = 0$, then $\mathrm{DAI}_m(\gamma_m, \delta_m) = \gamma_m + \delta_m D(0, \beta)$.

PROPOSITION 3: Denote by $\Phi(\epsilon)$ the standard normal cumulative distribution function. Then for each $\beta \in (0, 1)$ the real number $D(0, \beta)$ uniquely solves $D(0, \beta) = \{1 - \beta\Phi[D(0, \beta)]\}^{-1}\beta\Phi'[D(0, \beta)]$.

PROOF OF PROPOSITIONS 2 AND 3: From lemma 1,

$$\mathrm{DAI}_m(\gamma_m, \delta_m) = \sup_{\tau \in T} E_0\left[(1 - \beta) \sum_{r=0}^{\tau} \beta^r(x_{mr} - \psi_r) + \beta^{\tau+1}\mathrm{DAI}_m(\gamma_m, \delta_m)\right].$$

Supposing $\sigma_m = 0$, after one period's experience, the match parameter ξ_m is perfectly revealed. Therefore if $\xi_m \geqslant \mathrm{DAI}_m(\gamma_m, \delta_m)$, setting $\tau = \infty$ is optimal Otherwise $\tau = 0$ is optimal. Hence

$$\mathrm{DAI}_m(\gamma_m, \delta_m) = (1 - \beta)\gamma_m + \beta E_0\{\max[\xi_m, \mathrm{DAI}(\gamma_m, \delta_m)]\}.$$

Let $\Phi(\epsilon)$ denote the standard normal distribution function. Noting that beliefs about ξ_m are initially distributed $N(\gamma_m, \delta_m^2)$:

$$\begin{aligned}\mathrm{DAI}_m(\gamma_m, \delta_m) &= (1 - \beta)\gamma_m + \beta E_0\{\max[\gamma_m + \delta_m\epsilon, \mathrm{DAI}(\gamma_m, \delta_m)]\} \\ &= \gamma_m + \delta_m\beta E_0\left\{\max\left[\epsilon, \frac{\mathrm{DAI}(\gamma_m, \delta_m) - \gamma_m}{\delta_m}\right]\right\}.\end{aligned}$$

Let $x \equiv \delta_m^{-1}[\mathrm{DAI}_m(\gamma_m, \delta_m) - \gamma_m]$, subtract γ_m from both sides of the equation, and then divide through by δ_m to obtain the recursion

$$\begin{aligned} x &= \beta E_0[\max(\epsilon, x)] \\ &= \beta\left[x\Phi(x) + \int_x^{\infty} \epsilon\Phi'(\epsilon)d\epsilon\right] \\ &= \beta[x\Phi(x) + \Phi'(x)]. \end{aligned}$$

For each $\beta \in (0, 1)$ consider solutions to the equation above in x. From the definition of x, $\delta_m^{-1}[\mathrm{DAI}_m(\gamma_m, \delta_m) - \gamma_m]$ solves the recursion for all $\beta \in (0, 1)$. Setting $(\gamma_m, \delta_m) = (0, 1)$, proposition 3 is established on rearranging terms. Also, $\partial(\mathrm{LHS})/\partial x = 1$; $\partial(\mathrm{RHS})/\partial x = \beta\Phi(x) < 1$. Hence $D(0, \beta)$ is the unique solution to the equation above, implying $D(0, \beta) = \delta_m^{-1}[\mathrm{DAI}_m(\gamma_m, \delta_m) - \gamma_m]$, which, on rearrangement, yields proposition 2.

Proposition 4: $\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt}) = \gamma_{mt} + \delta_{mt}D(\alpha_m + \tau_{mt}, \beta)$

Proof of proposition 4: If $\sigma_m = 0$, the proof to propositions 2 and 3 applies. Accordingly, suppose $\sigma_m > 0$. From definition 2,

$$\begin{aligned} \mathrm{DAI}_m(\gamma_{mt}, \delta_{mt}) &= \sup_{\tau \geq t}\left\{\frac{E_t\left[\sum_{s=t}^{\tau} \beta^{s-t}(x_{ms} - \psi_s)\right]}{E_t\left(\sum_{s=t}^{\tau} \beta^{s-t}\right)}\right\} \\ &= \gamma_{mt} + \sup_{\tau \geq t}\left\{\frac{E_t\left[\sum_{s=t}^{\tau} \beta^{s-t}(x_{ms} - \psi_s - \gamma_{mt})\right]}{E_t\left(\sum_{s=t}^{\tau} \beta^{s-t}\right)}\right\} \end{aligned}$$

(adding and subtracting γ_{mt})

$$= \gamma_{mt} + \sup_{\tau \geq t}\left\{\frac{E_t\left[\sum_{s=t}^{\tau} \beta^{s-t}\delta_{mt}^{-1}(x_{ms} - \psi_s - \gamma_{mt})\right]}{[illegible]}\right\}\delta$$

(multiplying and dividing the b
tions (1) and (2) in the text.
+ $(\alpha_m + \tau_{mt})^{1/2}\epsilon_{mt}$. As of time
normally distributed $N(\gamma_{mt}, \delta^2_{mt}$
are normally distributed $N(0, 1)$.

$$D(\alpha_m + \tau_{mt}, \beta) = \sup_{\tau \geq t}\left\{\frac{E_t\left[\sum_{s=t}^{\tau} \beta \; [illegible]\right]}{E_t\left(\sum_{s=t}^{\tau} \beta^{s-t}\right)}\right\}$$

Substituting the left-hand side of this equation into the expression derived for $\mathrm{DAI}_m(\gamma_{mt}, \delta_{mt})$ above, one obtains the result for $\sigma_m > 0$.

Proposition 5: Let F denote the space of bounded continuous real-valued functions taking their domain on the space $(-\infty, \infty) \times (0, \infty)$. For all $f \in F$ define C, a mapping from F to itself, as

$$C[f(\gamma, \alpha)] = \beta \int_{-\infty}^{\infty} \max(0, (1 - \beta)\{\gamma + \epsilon[\alpha(\alpha + 1)]^{-1/2}\}$$

$$+ f\{\gamma + \epsilon[\alpha(\alpha + 1)]^{-1/2}, \alpha + 1\})d\Phi(\epsilon).$$

There exists a unique function $g(\gamma, \alpha) \in F$ satisfying $g(\gamma, \alpha) = C[g(\gamma, \alpha)]$. Moreover, defining $C_k(f)$ recursively as $C[C_{k-1}(f)]$, for all $f \in F$, $\|C_k(f) - g\| \leqslant (1 - \beta)^{-1}\|C_k(f) - C_{k-1}(f)\|$. Finally, for all $(\alpha, \beta) \in (0, \infty) \times (0, 1)$, $D(\alpha, \beta) = (1 - \beta)^{-1}\alpha^{1/2}g[-\alpha^{-1/2}D(\alpha, \beta), \alpha]$.

Proof of proposition 5: Note that F is complete and therefore, by Blackwell's (1965) sufficiency conditions, C is a contraction. Hence the contraction mapping theorem applies, simultaneously establishing the existence of a unique fixed point $g(\gamma, \alpha)$ and the inequality stated in the proposition (see Smart [1974] for details).

Now suppose $\sigma_m = 1$, which implies $\delta_m = \alpha_m^{1/2}$. Then, for all $y \in (-\infty, \infty)$, define

$$V(y, \gamma_m, \alpha_m) \equiv \sup_{\tau \in T} E_0\left[(1 - \beta)\sum_{r=0}^{\tau} \beta^r(x_{mr} - \psi_r) + \beta^{\tau+1}y\right]$$

$$g_1(\gamma_m, \alpha_m) \equiv \beta E_0 \max[0, V(0, \gamma_{m1}, \alpha_m + 1)]$$

One can establish $g(\gamma, \alpha) = g_1(\gamma, \alpha)$. To prove this claim, observe that

$$C[g_1(\gamma_m, \alpha_m)] = \beta \int_{-\infty}^{\infty} \max(0, (1 - \beta)\{\gamma + \epsilon[\alpha_m(\alpha_m + 1)]^{-1/2}\}$$
$$+ g_1\{\gamma_m + \epsilon[\alpha_m(\alpha_m + 1)]^{-1/2}, \alpha_m + 1\})d\Phi(\epsilon)$$

(from the definition of C)

$$= \beta E_0 \max[0, (1 - \beta)\gamma_{m1} + g_1(\gamma_{m1}, \alpha_m + 1)]$$

(using eq. [2])

$$= \beta E_0 \max\{0, (1 - \beta)\gamma_{m1}$$
$$+ \beta E_1 \max[0, V(0, \gamma_{m2}, \alpha_m + 2)]\}$$

(from the definition of g_1)

$$= \beta E_0 \max[0, V(0, \gamma_{m1}, \alpha_m + 1)]$$

(appealing to Bellman's [1957] principle of optimality)

$$\equiv g_1(\gamma_m, \alpha_m).$$

Thus g_1 is a fixed point for C on F. So from the uniqueness of g, the claimed identity between g_1 and g follows immediately. Hence, appealing to Bellman's principle once more, $V(0, \gamma_m, \alpha_m) = (1 - \beta)\gamma_m + g(\gamma_m, \alpha_m)$. Also, since $\sigma_m = 1$, the definition of $V(y, \gamma_m, \alpha_m)$ implies by lemma 1 that

$$V[\mathrm{DAI}_m(\gamma_m, \alpha_m^{-1/2}), \gamma_m, \alpha_m] = \mathrm{DAI}_m(\gamma_m, \alpha_m^{-1/2})$$
$$= \gamma_m + \alpha_m^{-1/2}D(\alpha_m, \beta),$$

the second line following from proposition 4. Setting $\gamma_m = -\alpha_m^{-1/2}D(\alpha_m, \beta)$, then,

$$
\begin{aligned}
0 &= \mathrm{DAI}_m(\gamma_m, \alpha_m^{-1/2}) \\
&= V[0, -\alpha_m^{1/2}D(\alpha_m, \beta), \alpha_m^{-1/2}] \\
&= -(1-\beta)\alpha_m^{-1/2}D(\alpha_m, \beta) + g[-\alpha_m^{1/2}D(\alpha_m, \beta), \alpha_m].
\end{aligned}
$$

Proposition 6: For all $(m, \tau) \in M \times T$

$$h_{m\tau} = \psi_{m\tau}[\alpha_m^{1/2}D(\alpha_m, \beta) - (\alpha_m + \tau)^{1/2}D(\alpha_m + \tau, \beta)]$$

$$\psi_{m1}(\rho) = \Phi[\rho\alpha_m^{1/2}(\alpha_m + 1)^{1/2}]$$

$$\psi_{m(\tau+1)}(\rho) = (1 - h_{m\tau})^{-1}\left\{\int_{-\infty}^{\infty} \psi_{m\tau}[\rho - \epsilon(\alpha_m + \tau)^{1/2}(\alpha_m + \tau + 1)^{1/2}]d\Phi(\epsilon) - h_{m\tau}\right\}.$$

Proof of proposition 6: From definition 6,

$$
\begin{aligned}
h_{m\tau} &= P\{\gamma_{mt} + \delta_{mt}D(\alpha_m + \tau, \beta) < \gamma_m + \delta_m D(\alpha_m, \beta) | \tau_{mt} = \tau\} \\
&= P\{\gamma_{mt} + \sigma_m(\alpha_m + \tau)^{-1/2}D(\alpha_m + \tau, \beta) < \gamma_m + \sigma_m\alpha_m^{-1/2}D(\alpha_m, \beta) | \tau_{mt} = \tau\}
\end{aligned}
$$

(noting from [2] that if $\tau_{mt} = \tau$ then $\delta_{mt} = \sigma_m[\alpha_m + \tau]^{-1/2}$)

$$= P\{\gamma_{mt} < \gamma_m + \sigma_m[\alpha_m^{-1/2}D(\alpha_m, \beta) - (\alpha_m + \tau)^{-1/2}D(\alpha_m + \tau, \beta)] | \tau_{mt} = \tau\}$$

(subtracting $\sigma_m[\alpha_m + \tau]^{-1/2}D[\alpha_m + \tau, \beta]$ from both sides of the inequality)

$$= \psi_{m\tau}[\alpha_m^{-1/2}D(\alpha_m, \beta) - (\alpha_m + \tau)^{-1/2}D(\alpha_m + \tau, \beta)]$$

(from def. 7). This proves the first equality in proposition 6. Also from definition 7,

$$
\begin{aligned}
\psi_{m1}(\rho) &= P\{\gamma_{mt} < \gamma_m + \rho\sigma_m | \tau_{mt} = 1\} \\
&= P\{\gamma_{m1} < \gamma_m + \rho\sigma_m | \tau_{m1} = 1\}
\end{aligned}
$$

(since from eq. [1] net returns $[x_{mt} - \psi_t]$ are identically and independently distributed over time)

$$= P\{[\delta_m^{-2}\gamma_m + \sigma_m^{-2}(x_{m1} - \psi_1)](\delta$$ [illegible] $$+ \rho\sigma\}$$

(from eq. [2])

$$= P\{[\alpha_m\gamma_m + (x_{m1} - \psi$$ [illegible]

(multiplying the numerator and [illegible]
noting $\alpha_m = \delta_m^{-2}\sigma_m^2$)

$$= P\{\sigma_m^{-1}(x_{m1} - \psi_1 - \gamma$$ [illegible]

(subtracting γ_m from both sides [illegible]
through by $\sigma_m^{-1}[1 + \alpha_m]$). But from (1), [illegible]
ϵ_{m1}. Beliefs about the first expression on the right-hand side [illegible]
bution $N(0, \alpha_m^{-1})$, while the second is independently distributed $N(0, 1)$.
Hence beliefs about the left-hand side are distributed $N(0, 1 + \alpha_m^{-1})$. Therefore, supposing ϵ is a standard normal random variate,

$$
\begin{aligned}
\psi_{m1} &= P\{\sigma_m^{-1}(x_{m1} - \psi_1 - \gamma_m) < \rho(1 + \alpha_m)\} \\
&= P\{(1 + \alpha_m^{-1})^{1/2}\epsilon < \rho(1 + \alpha_m)\} \\
&= P\{\epsilon < \rho(1 + \alpha_m)^{1/2}\alpha_m^{1/2}\} \\
&= \Phi[\rho\alpha_m^{1/2}(\alpha_m + 1)^{1/2}]
\end{aligned}
$$

Delhi

This proves the second equality in proposition 6. Finally if $d_{mt} = 1$ for all $t \leq \tau$, then from (2)

$$\begin{aligned}\gamma_{m(\tau+1)} &= [(\alpha_m + \tau)\gamma_{m\tau} + (x_{m\tau} - \psi_\tau)](\alpha_m + \tau + 1)^{-1} \\ &= (\alpha_m + \tau)(\alpha_m + \tau + 1)^{-1}\gamma_{m\tau} + (\alpha_m + \tau + 1)^{-1}(x_{m\tau} - \psi_\tau) \\ &= (\alpha_m + \tau)(\alpha_m + \tau + 1)^{-1}\gamma_{m\tau} + (\alpha_m + \tau + 1)^{-1}(\xi_m + \sigma_m\epsilon_{m\tau}).\end{aligned}$$

Since beliefs about ξ_m are distributed $N(\gamma_{m\tau}, \delta^2_{m\tau})$ when $t_{mt} = \tau$, and $\epsilon_{m\tau}$ is distributed $N(0, 1)$, $\xi_m + \sigma_m\epsilon_{mt} = \gamma_{m\tau} + (\delta^2_{m\tau} + \sigma^2_m)^{1/2}\epsilon$, where ϵ is distributed $N(0, 1)$. Hence

$$\begin{aligned}\gamma_{m(\tau+1)} &= (\alpha_m + \tau)(\alpha_m + \tau + 1)^{-1}\gamma_{m\tau} + (\alpha_m + \tau + 1)^{-1}[\gamma_{m\tau} + (\delta^2_{m\tau} + \sigma^2_m)^{1/2}\epsilon] \\ &= \gamma_{m\tau} + (\alpha_m + \tau + 1)^{-1}(\delta^2_{m\tau} + \sigma^2_m)^{1/2}\epsilon \\ &= \gamma_{m\tau} + \sigma_m(\alpha_m + \tau + 1)^{-1/2}(\alpha_m + \tau)^{-1/2}\epsilon,\end{aligned}$$

since $(\alpha_m + \tau) = \delta^{-2}_{m\tau}\sigma^2_m$ if $\tau_{mt} = \tau$. Appealing to definition 7 and substituting for $\gamma_{m(\tau+1)}$,

$$\begin{aligned}\psi_{m(\tau+1)}(\rho) &= P\{\gamma_{m(\tau+1)} < \gamma_m + \sigma_m\rho|\tau_{m(\tau+1)} = \tau + 1\} \\ &= P\{\gamma_{m\tau} < \gamma_m + \sigma_m[\rho - (\alpha_m + \tau + 1)^{-1/2} \\ &\quad \times (\alpha_m + \tau)^{-1/2}\epsilon]|\tau_{m\tau} = \tau, d^o_{m\tau} = 1\}.\end{aligned}$$

Using definition 6, observe that

$$\begin{aligned}P\{d^o_{m\tau} = 1|\tau_{m\tau} = \tau\} &= 1 - P\{d^o_{m\tau} = 0|\tau_{m\tau} = \tau\} \\ &= 1 - P\{\gamma_{m\tau} + \delta_{m\tau}D(\alpha_m + \tau, \beta) < \gamma_m \\ &\quad + \delta_m D(\alpha_m, \beta)|\tau_{m\tau} = \tau\} \\ &= 1 - h_{m\tau}.\end{aligned}$$

Combining the last equation with the expression obtained for $\psi_{m(\tau+1)}(\rho)$, one obtains

$$\begin{aligned}&(1 - h_{m\tau})\psi_{m(\tau+1)}(\rho) \\ &= P\{\gamma_{m\tau} < \gamma_m + \sigma_m[\rho - (\alpha_m + \tau + 1)^{-1/2}(\alpha_m + \tau)^{-1/2}\epsilon]|\tau_{m\tau} \\ &\qquad = \tau, d^o_{m\tau} = 1\}P\{d^o_{m\tau} = 1|\tau_{m\tau} = \tau\} \\ &= P\{\gamma_{m\tau} < \gamma_m + \sigma_m[\rho - (\alpha_m + \tau + 1)^{-1/2}(\alpha_m + \tau)^{-1/2}\epsilon], d^o_{m\tau} = 1|\tau_{m\tau} = \tau\} \\ &= P\{\delta_m D(\alpha_m, \beta) - \delta_{m\tau}D(\alpha_m + \tau, \beta) < (\gamma_{m\tau} - \gamma_m) \\ &\qquad < \sigma_m[\rho - (\alpha_m + \tau + 1)^{-1/2}(\alpha_m + \tau)^{-1/2}\epsilon]|\tau_{m\tau} = \tau\}\end{aligned}$$

(applying the DAI rule, given there exists a new job in the same occupation as m)

$$\begin{aligned}&= P\{(\gamma_{m\tau} - \gamma_m) < \sigma_m[\rho - (\alpha_m + \tau + 1)^{-1/2}(\alpha_m + \tau)^{-1/2}\epsilon]|\tau_{m\tau} = \tau\} \\ &\qquad - P\{\delta_m D(\alpha_m, \beta) - \delta_{m\tau}D(\alpha_m + \tau, \beta) < \gamma_{m\tau} - \gamma_m|\tau_{m\tau} = \tau\} \\ &= P\{\gamma_{m\tau} < \gamma_m + \sigma_m[\rho - (\alpha_m + \tau + 1)^{-1/2}(\alpha_m + \tau)^{-1/2}\epsilon]|\tau_{m\tau} = \tau\} - h_{m\tau}\end{aligned}$$

(using def. 6)

$$= \int_{-\infty}^{\infty} P\{\gamma_{m\tau} < \gamma_m + \sigma_m[\rho - (\alpha_m + \tau + 1)^{-1/2}(\alpha_m + \tau)^{-1/2}\epsilon]|\tau_{m\tau}$$

$$= \tau, \epsilon\}d\Phi(\epsilon) - h_{m\tau}$$

$$= \int_{-\infty}^{\infty} \psi_{m\tau}[\rho - (\alpha_m + \tau + 1)^{-1/2}(\alpha_m + \tau)^{-1/2}\epsilon]d\Phi(\epsilon) - h_{m\tau}$$

The last line follows from definition 7. Dividing through by $(1 - h_{m\tau})$ the third equation in proposition 6 is obtained, thus completing the proof

Appendix B

The procedure below describes how to compute the standard index Then the properties of the resulting approximation are discussed.

i) *Given* $(\alpha_j, \beta_k) \in [0, \infty] \times (0, 1)$, *recursively define the functions* $f_h(\gamma, \alpha_j - h)$, which map $[-\infty, \infty)$ into itself, as follows First, $f_0(\gamma, \alpha_j) = 0$. Second, $f_{h+1}(\gamma, \alpha_j - h - 1)$ is the cubic spline that knots together a finite number of coordinate pairs $[\gamma_i, f_{h+1}(\gamma_i, \alpha_j - h - 1)]$, the abscissa γ_i being drawn from a finite set $\{\gamma_i\}_{i\in I}$ and the ordinate $f_{h+1}(\gamma_i, \alpha_j - h - 1)$ being determined from $f_{h+1}(\gamma_i, \alpha_j - h - 1) = C[f_h(\gamma_i, \alpha_j - h)]$. (A cubic spline is a piecewise cubic polynomial that joins coordinate pairs so that the resulting function is twice differentiable everywhere.)

ii) *Given a finite set* $\{\alpha_j\}_{j\in J}$, *let* $\{\alpha_k\}_{k\in K}$ *be the linear transformation* $\{\alpha_k\}_{k\in K} = \{\alpha_j + 1\}_{j\in J}$.

iii) Observing from ii that $f_{h+1}(\gamma_i, \alpha_k - h - 1) = f_{h+1}(\gamma_i, \alpha_j - h)$, suppose that ν is the lowest counting number that, for all $(\gamma, \alpha) \in \{\gamma_i\}_{i\in I} \times \{\alpha_j\}_{j\in J}$, satisfies $|f_{\nu+1}(\gamma_i, \alpha_j - \nu) - f_\nu(\gamma_i, \alpha_j - \nu)| \leq 0.001(1 - \beta_k)$

iv) Finally, let D_{jk} denote the (unique) solution to the equation $D_{jk} = (1 - \beta_k)^{-1}(\alpha_j - \nu)^{1/2}f_\nu[(\alpha_j - \nu)^{-1/2}D_{jk}, \alpha_j - \nu]$ The number $D(\alpha_j - \nu, \beta_k)$ was accordingly approximated by D_{jk}. The theoretical justification for this approximation comes from proposition 5. if $\|C[f_\nu(\gamma, \alpha)] - f_\nu(\gamma, \alpha)\| \leq 0.001(1 - \beta)$, then $\|g(\gamma, \alpha) - f_\nu(\gamma, \alpha)\| \leq 0.001$. Upon substituting $f_\nu(\gamma, \alpha)$ for $g(\gamma, \alpha)$ in the equation $D(\alpha - \nu, \beta) = (1 - \beta)^{-1}(\alpha$ [illegible]
the approximation proposed abo[illegible]

This algorithm is subject to tw[illegible]
$\| \ \|$ is applied only to the finite se[illegible]
set $[-\infty, \infty] \times [0, \infty]$. Second, the [illegible]
determine $D(\alpha, \beta_k)$, rather than g[illegible]
apparently, be cheaply controlle[illegible]
than 0.95. At high levels of β, t[illegible]
increased to satisfy the convergence [illegible]

Table B1 reports computed values of D[illegible]
β). The numbers down the left side of the table indicate values of α [illegible]
..., 17.1}, while those across the top refer to values of $\beta \in \{0.1,$ [illegible] $, 0.95\}$

Appendix C

The following procedure describes how to compute the sequence of hazards $\{h_t(\alpha, \beta)\}_{t=1}^{5}$ for different values of α and β

i) For each β_k compute a cubic spline for $D(\alpha, \beta_k)$ by knotting together the points determined in Appendix B

ii) Appealing to the second equation of proposition 6, compute $\psi_{m1}(\rho)$ directly. Hence, h_{m1} is determined from the first equation in proposition 6 by

TABLE B1

$D(\alpha_m, \beta)$

	.10	.20	30	.40	50	60	70	80	90	.95
.0	.0420	0890	1422	.2036	2760	3644	.4777	.6360	.9015	1.1589
1	0396	0840	.1344	1926	.2614	3457	4538	.6049	8562	1.0925
.3	.0365	0775	.1243	1787	.2433	3229	.4260	5712	.8149	1.0458
7	0319	0681	.1096	1582	2165	.2890	3839	.5193	7512	9722
1.1	.0287	0615	0992	1436	.1972	.2642	.3525	4809	.7021	.9144
1.3	0275	0588	0950	1377	.1893	2540	3402	4648	.6814	.8898
1.7	.0253	.0543	0879	1277	.1760	.2368	3185	.4371	.6455	.8468
2.1	0236	0508	.0823	.1197	.1652	.2231	3005	4143	.6144	.8085
3 1	.0205	.0442	.0718	1047	.1450	1968	2664	.3704	5539	7323
4.1	0183	0396	0644	.0941	.1306	1780	.2420	3370	5077	.6748
5.1	.0167	.0361	.0589	.0862	.1198	1638	2236	.3122	.4702	.6269
6.1	.0155	.0334	.0545	.0799	.1113	.1527	.2090	2926	.4421	.5856
7 1	.0144	.0312	.0509	0747	1043	1434	1970	2766	.4189	.5540
8.1	.0136	0293	0479	0703	.0983	1354	1866	.2627	.4004	5282
9 1	.0128	0278	0453	.0667	.0931	.1283	.1771	.2500	.3860	.5075
10.1	.0122	0260	0430	.0634	.0885	1220	1686	.2385	.3719	4900
11 1	.0116	0248	0410	.0605	.0844	.1162	.1606	.2296	3582	4736
12 1	.0111	.0237	.0394	.0578	.0806	1108	1534	.2202	.3458	.4572
13.1	0107	.0227	.0377	.0553	.0770	.1066	1476	.2113	3336	4424
14.1	0103	.0218	.0363	0532	.0742	1023	1415	2028	.3214	.4277
15.1	0099	.0210	.0350	0512	.0713	.0969	1358	1947	.3092	4123
16.1	0096	.0203	.0338	0494	.0688	.0932	.1304	.1869	.2969	3960
17.1	.0093	0196	0327	.0478	.0664	0900	1257	.1795	2847	.3791

using step i to evaluate $\psi_{m1}(\rho)$ at the number $[\alpha^{1/2}D(\alpha, \beta_k) - (\alpha + 1)^{1/2}D(\alpha + 1, \beta_k)]$.

iii) Substitute $\psi_{m1}(\rho)$ and h_{m1} into the third equation of proposition 6. Then compute $\psi_{m2}(\rho)$. Using step i again, evaluate this distribution at the number $[\alpha^{1/2}D(\alpha, \beta_k) - (\alpha + 2)^{1/2}D(\alpha + \tau, \beta_k)]$ to determine h_{m2}. Continuing in this fashion, h_{m3}, h_{m4}, and h_{m5} are successively determined.

Table C1 reports, for 230 coordinate pairs (α_m, β), the computed sequences $\{h_{m\tau}(\alpha_m, \beta)\}_{\tau=1}^{5}$. The numbers down the left side of the table indicate values of $\alpha_m \in [0.1, \ldots, 17.1]$, while those across the top refer to values of $\beta \in [0.1, \ldots, 0.95]$.

Appendix D

Not all the 822 life histories in the subsample of whites were adequately documented, and a number of conditions were imposed that individual data were required to satisfy for inclusion in the analysis.

i) Information about full-time activities had to be continuously recorded from the time of entry into the labor force through to January 1969. Full-time activities comprised full-time education, full-time employment, and full-time other, all of which were explicitly coded (loss 200).

ii) No overlaps in full-time activities were permitted. Overlaps can occur because of miscoded spell dates, multiple full-time jobs being held simultaneously, or, in some cases, the classification of seasonal work (brought about, e.g., by a schoolteacher's taking a summer job). It was difficult to distinguish these three cases without access to the questionnaires, and so respondents with any overlapping full-time employment spells were dropped (loss 141).

iii) Respondents with missing data on spell dates and relevant variable codes were dropped (loss 5), as were those whose records claimed a spell finished before it started (loss 18). With respect to conditions i–iii, for full-time spells, spell beginning and spell ending dates are available in two forms, a continuous numeric code and a month/year code. When cases failed to satisfy all of the conditions above using the numeric codes, an attempt was made to salvage them using the year/month [illegible]
so, after netting out two who pursue[illegible]
work history, this left the 167 wh[illegible]
study.

References

Bartel, Ann P. "The Migration D[illegible]
A.E.R. 69 (December 1979): 7[illegible]

Bartel, Ann P., and Borjas, George J[illegible]
Empirical Analysis." In *Studies in Labor* [illegible]
Chicago: Univ. Chicago Press (for N.B.E.R.), 1981.

Becker, Gary S. "Investment in Human Capital: A Theoretical Analysis." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 9–49.

——— *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education.* 2d ed. Chicago: Univ. Chicago Press (for N.B.E.R.), 1975.

Bellman, Richard E. *Dynamic Programming.* Princeton, N.J.: Princeton Univ. Press, 1957.

Blackwell, David. "Discounted Dynamic Programming." *Annals Math. Statis.* 36 (February 1965): 226–35.

Borjas, George J., and Rosen, Sherwin. "Income Prospects and Mobility of

TABLE C1

$\{h_{m\tau}(\alpha_m, \beta)\}_{\tau=1}^{5}$

	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
.1	.513	.527	.544	.562	.584	.610	.642	.685	.750	.804
	.071	.075	.079	.084	.089	.097	.107	.123	.151	.181
	.029	.030	.032	.034	.037	.040	.045	.053	.067	.083
	.015	.016	.017	.019	.020	.022	.025	.030	.039	.048
	.010	.010	.011	.012	.013	.014	.016	.019	.025	.032
.3	.511	.522	.536	.551	.569	.591	.618	.654	.711	.760
	.114	.119	.124	.130	.137	.147	.160	.179	.213	.248
	.050	.052	.054	.057	.061	.066	.073	.083	.103	.123
	.028	.029	.030	.032	.035	.038	.042	.048	.061	.074
	.018	.018	.019	.021	.022	.024	.027	.031	.040	.050
.5	.508	.516	.526	.538	.551	.567	.588	.616	.661	.701
	.138	.142	.147	.153	.161	.170	.182	.201	.234	.268
	.063	.065	.068	.071	.075	.080	.088	.098	.118	.139
	.036	.037	.039	.041	.044	.047	.051	.058	.072	.086
	.023	.024	.025	.027	.028	.031	.034	.039	.048	.059
.7	.508	.517	.528	.540	.554	.571	.593	.622	.670	.712
	.154	.159	.164	.170	.178	.188	.201	.220	.255	.290
	.073	.076	.079	.082	.087	.092	.100	.112	.133	.156
	.043	.044	.046	.048	.051	.055	.060	.068	.082	.098
	.028	.029	.030	.032	.034	.036	.040	.045	.056	.068
.9	.508	.517	.527	.538	.552	.569	.590	.619	.666	.708
	.166	.170	.176	.182	.190	.200	.213	.232	.267	.301
	.081	.084	.087	.091	.095	.101	.109	.121	.143	.166
	.048	.050	.052	.054	.057	.061	.066	.074	.090	.106
	.032	.033	.034	.036	.038	.041	.044	.050	.061	.074
1.1	.507	.514	.523	.533	.545	.559	.578	.603	.645	.683
	.175	.179	.184	.190	.198	.207	.219	.237	.270	.303
	.088	.090	.093	.097	.102	.107	.115	.127	.149	.171
	.053	.054	.056	.059	.062	.066	.071	.079	.094	.111
	.035	.036	.038	.039	.041	.044	.048	.054	.065	.078

1.5	.506	.512	.520	529	539	551	.568	591	.630	.665
	.188	.192	197	202	.209	.218	.230	247	.277	.308
	.098	.101	104	.107	.112	117	.125	.136	.158	179
	.060	062	.064	067	069	073	079	.087	102	.119
	.041	.042	.044	045	047	.050	.054	060	.071	.085
2 1	.505	510	516	524	532	.543	.557	.576	610	.641
	201	204	209	214	.220	228	238	254	.281	.309
	.109	112	115	.118	.122	127	.134	145	.166	.186
	069	071	.073	075	078	082	087	.094	.109	.126
	.048	049	050	052	054	057	061	067	.078	091
3 1	.504	508	513	518	525	534	.545	.562	.590	616
	213	.216	220	224	229	236	245	258	.283	307
	122	124	126	129	133	137	144	153	171	192
	079	081	083	085	.087	091	096	103	116	131
	056	057	059	060	062	065	068	074	084	.096
4 1	503	506		515	521	.528	537	551	577	.600
	221	224		230	235	240	248	.259	280	.304
	129	[illegible]		136	139	144	.149	158	173	.191
	086	[illegible]		091	094	097	101	108	119	.133
	062			066	068	070	074	.079	088	099
5 1	.503			513	518	.524	532	544	565	589
	226			234	238	243	250	260	278	.298
	135			141	144	148	153	.161	174	.190
	091			096	098	101	105	.111	121	133
	066			070	072	074	.078	083	091	.100
6 1	502			511	516	521	528	538	557	576
	230			237	241	245	251	260	.275	.293
	139			145	147	151	156	.163	173	187
	095			099	.102	105	108	.114	123	133
	070			073	075	077	081	085	093	.101
7 1	502			510	514	519	525	535	550	567
	233			240	243	248	253	262	273	.289
	143			148	150	154	158	165	175	.186
	098			102	104	107	111	116	125	134
	073			076	078	080	083	087	095	103

TABLE C1 (*Continued*)

$\{h_{mi}(\alpha_m, \beta_i)\}_{i=1}^{5}$

	.10	20	30	40	.50	60	.70	80	.90	.95
8.1	502	504	.506	509	.513	517	523	532	.543	.559
	.237	.239	240	243	246	250	.255	.263	274	.286
	.146	147	149	.150	153	156	.161	.166	.177	.186
	.101	102	.103	105	.107	.109	.113	118	.126	.135
	.075	076	077	078	080	.082	085	089	097	.104
9 1	.502	.504	506	508	.512	516	522	.530	.542	.554
	240	241	243	245	.248	.252	.257	.263	.275	.286
	.149	150	151	.153	156	.159	163	168	.178	188
	103	.104	.106	107	.109	111	115	.120	.128	.136
	077	078	079	080	082	.084	.087	.091	.098	.105
10.1	.501	503	505	.508	.511	515	520	526	.540	.551
	.243	.244	246	248	.250	254	259	265	.276	.287
	.152	.153	.155	156	159	.161	.165	.171	.181	.190
	106	.107	.108	109	.111	.114	.117	122	.131	.138
	.079	.080	081	082	083	.086	.088	093	100	.107
11 1	.501	503	.505	507	.510	514	519	.526	.538	.549
	245	247	.248	250	.253	.255	.259	.267	.277	.287
	156	.157	.158	.160	.162	.164	168	.174	.184	.192
	.109	110	111	112	114	.117	119	124	.133	.141
	.081	.082	083	084	.085	088	090	.095	103	110
12.1	.501	503	505	507	.509	512	.517	.525	.536	.546
	.247	.249	250	252	254	.257	.261	268	.278	.287
	160	.161	162	163	165	169	172	.177	.188	.196
	112	.113	114	115	117	119	122	127	.136	.145
	084	084	085	.086	088	.090	092	097	105	113

13.1	.501	503	504	506	508	512	.517	524	.536	545
	.249	250	.251	253	255	259	262	268	.279	288
	.164	.165	166	167	169	172	175	181	192	201
	115	116	117	118	.120	122	125	.130	140	.149
	.087	.088	089	090	.091	.093	.096	100	.109	118
14 1	501	502	.504	506	508	.513	516	523	.535	.545
	.250	251	252	253	255	.258	262	.268	.279	.289
	168	169	170	171	.173	175	.179	184	195	.206
	120	121	121	123	124	126	129	134	144	.154
	090	091	.092	.093	094	096	.098	102	112	.121
15.1	501	502	504	505	.508	.511	515	522	535	.546
	250	251	252	.253	255	257	261	267	278	.289
	173	174	175	176	178	180	183	.189	.199	211
	124	125	125	127	128	130	133	137	.148	159
	093	094	095	096	097	098	101	105	114	125
16.1	501	502		505	.507	510	514	.521	.534	.546
	248	249		251	253	255	258	264	275	.287
	176	177		179	181	183	186	190	202	214
	128	12		31	132	134	.136	141	151	163
	096	09		[illegible]	100	101	104	.108	.115	.128
17.1	.501	50		5	.507	.509	513	520	533	.547
	246	21		[illegible]	251	253	256	.260	272	284
	179	17		2	183	185	188	192	203	215
	132	1		34	136	137	140	144	.152	166
	100	1		12	103	104	107	110	.118	.127

Younger Men." Report no. 7920. Chicago: Univ. Chicago, Center Math. Studies Bus. and Econ., May 1979.

Gittins, J. C. "Bandit Processes and Dynamic Allocation Indices." *J. Royal Statis Soc*, ser. B, 41, no. 2 (1979): 148–77.

Gittins, J. C., and Jones, D. M. "A Dynamic Allocation Index for the Sequential Design of Experiments." In *Progress in Statistics*, edited by Joseph M. Gani. Amsterdam: North-Holland, 1974.

Heckman, James J., and Borjas, George J. "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence." *Economica* 47 (August 1980): 247–83.

Heckman, James J., and Singer, Burton. "The Identification Problem in Econometric Models for Duration Data." In *Advances in Econometrics*, edited by Werner Hildenbrand. London: Cambridge Univ. Press, 1982.

———. "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica* 52 (March 1984): 271–320.

Jovanovic, Boyan. "Job Matching and the Theory of Turnover." *J.P.E.* 87, no. 5, pt. 1 (October 1979): 972–90.

Laird, Nan. "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution." *J. American Statis. Assoc.* 73 (December 1978): 805–11.

Lippman, Steven A., and McCall, John J. "The Economics of Job Search: A Survey: Part I." *Econ. Inquiry* 14 (June 1976): 155–89.

Marshall, Alfred. *Principles of Economics*. 8th ed. New York: Macmillan, 1948

Miller, Robert A. "Job Specific Capital and Labor Mobility." Ph.D. dissertation, Univ. Chicago, June 1982.

Mincer, Jacob. "On-the-Job Training: Costs, Returns, and Some Implications." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 50–79.

Mincer, Jacob, and Jovanovic, Boyan. "Labor Mobility and Wages." In *Studies in Labor Markets*, edited by Sherwin Rosen. Chicago: Univ. Chicago Press (for N.B.E.R.), 1981.

Oi, Walter Y. "Labor as a Quasi-fixed Factor." *J.P.E.* 70, no. 5, pt. 2 (October 1962): 538–55.

Ornstein, Michael D. *Entry into the American Labor Force*. New York: Academic Press, 1976

Smart, D. R. *Fixed Point Theorems*. London: Cambridge Univ. Press, 1974.

Smith, Adam. *An Inquiry into the Nature and Causes of the Wealth of Nations*. 1776. Reprint ed. New York: Random House, 1937.

Viscusi, W. Kip. "Job Hazards and Worker Quit Rates: An Analysis of Adaptive Worker Behavior." *Internat. Econ. Rev.* 20 (February 1979): 29–58.

Weitzman, Martin L. "Optimal Search for the Best Alternative." *Econometrica* 47 (May 1979): 641–54.

Zellner, Arnold. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley, 1971.

Tenancy Choice in a Competitive Framework with Transactions Costs

Lee J. Alston
Williams College

Samar K. Datta and Jeffrey B. Nugent
University of Southern California

The choice between wage labor and sharecrop contracts is analyzed in a model that avoids both the pitfalls of previous models—inefficiency of sharecropping and indeterminacy—by introducing transactions costs with appropriate nonlinearities. Plantation data from the American South provide empirical support for the hypothesized nonlinearity in transactions costs and simultaneous determination of direct supervision and contractual choice.

Although transactions costs-
forcement costs associated w
siderable attention as explai
firm (Alchian and Demsetz 1
Meckling 1976), the levels a
liamson 1967; Calvo and We
(Ben-Porath 1980; Pollak 1982
tory (North 1981), somewhat surprising

We express our appreciation to Phillip Trevezas and Anabella Lardé de Palomo for research assistance, and to Bernardo Gonzales-Arechiga, Robert Higgs, Abdelhamid Mahbob, Morton Schapiro, Anup Sinha, and an anonymous referee for many useful suggestions.

[*Journal of Political Economy*, 1984, vol. 92, no. 6]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9206-0005$01.50

ations have not played such a prominent role in the literature on contractual choice in agrarian societies.[1]

In view of the particular difficulty that has arisen in explaining the existence of sharecropping and its pattern of variation from country to country, region to region, and crop to crop, we demonstrate that if transactions costs—typically far from negligible in the agrarian context (Higgs 1974; Alston 1981; Alston and Higgs 1982)—are incorporated into an otherwise standard neoclassical model, the existence of sharecropping can indeed be explained. Moreover, this can be done without resort to the ad hoc allocation of monopolistic powers across the different agents or to the introduction of interlinking with credit. Since monopolistic powers require some transactions costs to maintain, and since interlinking can be considered indirect supervision but would also require direct supervision in order to rule out access to the regular (noninterlinked) markets, the transactions cost explanation may well be considered more fundamental than the other explanations.[2]

Sharecropping literature has foundered on two alternative views about how the labor inputs of sharecroppers are determined. The traditional and still probably the dominant view is that the choice with respect to the level of labor inputs is the exclusive prerogative of the sharecrop tenant himself. Alternatively, the neoclassical view is that labor inputs are determined through the interaction of both parties, landowners and sharecroppers. Neither view, however, is satisfactory. The shortcoming of the neoclassical approach is that choice among alternative contractual forms becomes indeterminate; all forms are equally efficient. Since in the real world the pattern of contractual choice is much more systematic, with certain forms being dominant in certain regions, time periods, or crops, the neoclassical view has no explanatory power without transactions costs.[3] The problem of the traditional approach is that, even if it is couched in terms of a general equilibrium framework (e.g., as in Bardhan and Srinivasan [1971]), there is no way of conveying to the tenant the scarcity value of land if he is allowed to determine unilaterally the labor input per unit of land on sharecropped land. As a result, the tenant tends to supply infinitely small amounts of labor to a given amount of land (or, equivalently, to demand an infinitely large quantity of land for a given quantity of labor), yielding an "equilibrium" solution in which the marginal productivity of sharecropped land drops to zero and there is no reason for sharecropping to exist (Newbery 1974; Bagchi 1975; Reid 1976).

---

[1] The exceptions include Cheung (1969), Higgs (1974), Newbery (1975), Stiglitz (1975), Roumasset (1976), Lucas (1979), Newbery and Stiglitz (1979), Roumasset and Uy (1980), Alston (1981), Kotwal (1981*a*, 1981*b*), Alston and Higgs (1982), and Hayami and Kikuchi (1982). However, none of these authors has formally applied supervision costs to determine a unique solution *with* sharecropping.

[2] See, e.g., Braverman and Stiglitz (1982). With reference to their model, when the landlord wants either to encourage (through subsidization) or to penalize (through taxation) the tenant's use of credit, supervision would have to be applied to block the tenant's access to the credit market.

[3] The closest thing to an exception is the paper by Pant (1983). Pant's model, however, departs substantially from the standard neoclassical model in that it utilizes production functions with variable returns to scale and highly restrictive assumptions about the substitutability and marketability of different types of labor

Although an ideal theoretical model of contractual choice would incorporate all types of transactions costs, this paper deals only with supervision costs. This does not strain the credibility of our model, because empirically supervision costs have been the most powerful explanatory variable among transactions costs (Higgs 1974; Alston 1981; Alston and Higgs 1982). Further, in the empirical section of this paper, we are able to control for risk and for negotiation and enforcement costs. In our model the worker-sharecropper can be given the right to choose his labor input (and hence to shirk) but is constrained from doing so by the penalty imposed if he is caught shirking by the supervisor(s).[4] Through supervision and the threat of penalties, the scarcity value of land is communicated to the tenant, thereby constraining not only his labor supply (for given land) but also his demand for land (with given labor). As long as there are appropriate nonlinearities in these supervision costs (assumed to be the responsibility of the landowner), the resulting solution avoids not only the inefficiency problem of the traditional approach but also the indeterminacy problem of the neoclassical approach.

In the Appendix we present a sim cost
of the choice between wag
factors labor and land as in
strate that if supervision cos
cluded, the model yields solu
ties.[5] Without transactions

---

[4] For reasons given by Hayami and Kik
to be collected by the landlord but rather as a
discharge for poor performance. The assumption that the penalty varies with the extent of shirking can be thought of as affecting the character of the statement of discharge—e.g., a "bad-conduct" discharge for poor performance and "very-bad conduct" discharge for very poor performance. The fact that the penalty the worker-tenant pays for being detected shirking (or in this case moonlighting) does not accrue to the landowner is attributed to the static character of the model. In a more dynamic model, the disrepute the tenant earns from the penalty could be reflected in a lower wage rate or a higher crop-sharing rate for the landlord in the subsequent period.

[5] Nonsubstitutability and nonmarketability of different labor types in Pant (1983) could be interpreted as involving differential transactions costs in their employment. Moreover, his variable-to-scale production function serves the purpose of providing the appropriate nonlinearities for achieving an interior solution.

either into the neoclassical approach with indeterminacy but no inefficiency or into determinacy with inefficiency.

Hypotheses and Tests

In a generalized model of contractual mix, choice of contract depends on how transactions costs—negotiation, supervision, and enforcement costs—and risk vary across contracts.

In considering the three agricultural contract forms—wage, share, and fixed rent—there is little reason to believe that negotiation costs vary across contracts (Alston 1981). On the other hand, there is every reason to believe that supervision costs vary by type of contract. In particular, supervision costs tend to decrease the more closely effort is linked with reward. For this reason, supervision costs are greatest under wage contracts, since the reward—the wage—is based on the amount of time a worker spends rather than on his output. On the other hand, supervision costs are least under a fixed-rent contract because the landlord receives his payment for land regardless of the length and intensity of work effort, provided either that the payment is prior to the harvest or that the work effort is certain to yield a sufficient harvest to pay the landlord. Since workers under sharecropping receive only a share of their marginal products, sharecrop contracts occupy an intermediate ground in the supervision cost spectrum between wage contracts at the high end and fixed-rent contracts at the low end. This simple ordering of contractual arrangements, however, might not hold if other factors, such as the underreporting of output or land mismanagement, were considered.

In addition to supervising labor work effort, the landlord has an incentive to monitor the use of any other inputs that he supplies to the production process (Alston and Higgs 1982). Careless or excessive use of the landlord's work stock, equipment, or land would result in the depreciation of such assets. The more inputs and the more valuable are the inputs that the landlord supplies, the greater is the incentive to monitor. The marginal cost of monitoring inputs, however, tends to decrease as the number of inputs supplied increases. For example, the marginal cost of supervising labor work effort may be small if the landlord already monitors the use of his work stock by the worker.

In addition to these economies of scope, there may exist economies of scale in supervision—in terms of the number of employees monitored—up to the point where the supervisory capacity of the owner is reached. When another supervisor is hired, there is a discontinuous jump in supervision costs, after which marginal supervisory costs again most likely fall until the capacity of that supervisor is fully utilized. Other things being equal, the marginal costs of supervision are not as low for hired supervisors as for owner-supervisors because of the need to monitor nonowner supervisors. (Such costs, however, might be reduced if the hired supervisors were relatives.)

One might wonder why landlords could not dispense with the direct supervision of labor effort by indirectly monitoring it through the measurement of output.[6] For example, if output were to fall below some mutually agreed-on amount, labor could be penalized by the nonrenewal of contract or by assignment of a smaller share—provided the reason for low output was the fault of insufficient labor effort and not of some factor beyond the control of the worker (weather, insect infestation inter alia).

Although this possibility is theoretically appealing in a world free of transactions costs, several practical difficulties limit the degree to which landlords can monitor labor input by monitoring crop output. First, as already noted, landlords have the incentive to monitor the use of landlord-owned inputs whose value could be damaged by excessive or careless use by the worker, such as work stock, fertilizer, and especially land. Second, the effects of and hence the importance of some inputs and labor operations (especially fertilizer, depth of plowing, proper terracing, pruning of trees, and fencing) are not limited to, and indeed may not even be felt on, current output. Rather, the benefits may be realized primarily on future output. Hence, current output will be a poor indicator of current inputs. Third, even if it should be clear through output that a certain input was not supplied, chiefly that the worker did not work, it would be impossible for the landowner to know whether this was because the worker was ill and incapacitated by [illegible] (e.g., an epidemic) or becau [illegible] important, because of the [illegible] duction, landowners will hav [illegible] for shortfalls in production [illegible]

For all these reasons, th [illegible] monitor inputs directly. Nev [illegible] labor input would be prohibit [illegible] also rely somewhat on monitoring [illegible] output. Certainly workers and landlords have some mutually agreed-on rough set of expectations of output given diligent labor input. These expectations partly determine the fixed rent and the percentage the landlord receives under a share contract. If output is below the landlord's expectation, a worker's contract will most likely not be renewed—at least not under the same terms.

---

[6] We are grateful to an anonymous referee for raising this issue.

Along with supervision costs, enforcement costs vary across contractual form. Enforcement costs pertain to the costs of ensuring that workers honor the length of the contract. The greater the opportunity cost of breaking a contract, the less likely it will occur. Opportunity costs increase with the expectation of a larger payment at harvesttime, and decrease with the opportunity for off-farm earnings (Alston 1981). Assuming labor markets are competitive, labor's expected income will be greater under fixed-rent contracts than sharecrop contracts, because a fixed-renter requires lower costs on the part of the landlord (less supervision) than a sharecropper and receives a risk premium for bearing more risk. Hence fixed-renters are more likely to stay through harvest than sharecroppers. Similarly, sharecrop contracts secure labor better than wage contracts do because the opportunity cost of leaving (expected payment at harvest) exceeds that of wage hands who are paid on a time basis, bear no risk, and require considerable supervision.

Supervision costs take on many different forms, many of which are difficult to measure. The fact that our explanation hinges so heavily on considerations that are not easily quantified renders it vulnerable to the charge that it may be nonoperational, no matter how convenient it might be theoretically. Indeed, it is true that most sources of agricultural data have not even sought to measure the various kinds of transactions costs mentioned above.

Nevertheless, there are a few exceptions. One such exception was the data collected from large plantations in the Georgia "cotton belt" of the American South as part of a special survey undertaken by the U.S. Census Bureau in 1911 (Brooks 1914). Although the sample was small, this data source is extremely useful for testing hypotheses concerning supervision costs for the following reasons:

a) Risk and enforcement costs were controlled for by the homogeneous nature of the region surveyed. There is little reason to believe that risk and off-farm earnings opportunities varied over the area sampled.

b) Information on the number of supervisors was collected and, moreover, this information distinguishes between the numbers of owner-supervisors, who may be considered to be exogenous, and of nonowner or hired supervisors, who may be considered a decision variable of the landowner.

c) Data were collected on variables that permit the testing of hypotheses concerning the existence of economies of scope and scale on supervision and nonlinearities in the latter.

d) All three forms of labor contracts—fixed-rent, sharecrop, and wage—were present, and the percentages of each varied considerably from one plantation to another.

Essentially there are three ways for landlords to deal with labor supervision costs: (1) supervise labor directly, (2) hire labor supervisors, or (3) provide labor with an incentive to self-supervise by altering the contract mix. Consistent with the hypothesized simultaneity in decisions with respect to direct supervision and contractual choice, we test for the influence of several of the factors mentioned on such decisions by estimating the parameters in the following equations by two-stage least squares:

$$PF = \alpha_0 + \alpha_1 \frac{S}{A} + \alpha_2 PCA + \alpha_3 POWS, \tag{1}$$

$$PS = \beta_0 + \beta_1 \frac{S}{A} + \beta_2 PCA + \beta_3 POWS, \tag{2}$$

$$\frac{S}{A} = \gamma_0 + \gamma_1 PF + \gamma_2 PS + \gamma_3 \frac{OS}{A}, \tag{3}$$

where PF = fixed-rent contracts as a percentage of all contracts, PS = sharecrop contracts as a percentage of all contracts,[7] S/A = number of nonowner supervisors per 100 improved acres, PCA = cotton acreage as a percentage of all improved acreage, $POWS$ = owner's work stock as a percentage of all work stock, and OS/A = number of owner-supervisors per 100 improved acres.

Beginning with equation (1) for PF, the greater the number of hired supervisors, the lower would be the incentive to induce self-supervision through fixed-rent contracts; hence $\alpha_1 < 0$. Since cotton is a relatively labor-intensive crop (or at least was in the early twentieth century), the greater the percentage [illegible] acreage in cotton the higher would be the supe[illegible] benefits of inducing self-su[illegible] *the aforementioned econom*[illegible] the share of the owner's wor[illegible] incentive the owner has to [illegible] instead of inducing self-su[illegible] *hence* $\alpha_3 < 0$. *In equation* [illegible] mediate between fixed-rent and wage [illegible] costs are concerned, the signs of β_1, β_2, and β_3 are ambiguous. In equation (3), because of the supervision-reducing effects of share and

---

[7] Although it may be more conventional to measure the contractual mix in terms of percentages of land area, in this case an almost perfect positive correlation between land and number of labor contracts allows us to use percentages of labor contracts to *represent contractual mix.* Moreover, since the observations were large plantations, most of which had all three kinds of contracts but in varying proportions, PF and PS are not zero-one dummy variables (which would require a different estimation procedure), but rather continuous variables.

TABLE 1

MEANS AND REGRESSION RESULTS

	EQUATION			
RIGHT-HAND-SIDE VARIABLES	(1) *PF*	(2) *PS*	(3) *S/A*	MEANS
Intercept	96	−.07	.29	
	(7.08)	(.24)	(3.80)	
S/A	− 76	−.68		.094
	(1.62)	(.64)		
PCA	.20	82		474
	(1.07)	(1.98)		
POWS	−.72	.32		.864
	(4 99)	(98)		
PF			− 25	.168
			(2.10)	
PS			−.24	.529
			(2 32)	
OS/A			− 41	.059
			(2.06)	
R^2	.71	.25	38	

SOURCES.—Special investigation of Georgia plantations conducted by the U.S. Census Bureau in 1911. The plantation schedules can be found in the Rare Books and Manuscripts Department of the University of Georgia Library, Athens, Georgia.

NOTE.—Figures in parentheses are *t*-statistics. $N = 22$.

fixed-rent contracts relative to wage contracts, one would expect $\gamma_1, \gamma_2 < 0$. Finally, because the presence of an owner-supervisor would reduce the need for hired supervisors, one would also expect $\gamma_3 < 0$. While in fact *PF* and *PS* might be affected somewhat by *OS/A*, our inclusion of *OS/A* only in equation (3) where its effect would be expected to be stronger greatly facilitates identification of the complete model.

The 1911 data from the Georgia cotton belt are used to estimate the equations of the model via two-stage least squares. The results are given in table 1. As hypothesized, the values of α_1 and α_3 in the *PF* equation are negative and significant at the 5 percent level (in one-tailed tests), and the estimate of α_2 is positive but not significant. Also, the results of the *PS* equation show that $\beta_2 > 0$.

The most important results are those of equation (3) for *S/A*. As hypothesized γ_1, γ_2, and γ_3 are significantly negative. The negative γ_2 is particularly relevant inasmuch as it is indicative of nonlinearities in supervision costs per improved acreage of land with respect to area under sharecropping. The more land under fixed rent and sharecropping, the lower the need for nonowner supervisors per 100 improved acres. Likewise, the negative value of γ_3 indicates that the presence of an owner-supervisor (*OS/A*) significantly reduces the need for nonowner supervisors.

In general, therefore, the results are consistent with (1) the hypothesis that fixed-rent and sharecrop contracts reduce the need for direct supervision of labor, (2) the hypothesis that given the number of owner-supervisors per 100 improved acres the number of hired supervisors and contractual mix would be determined simultaneously, and (3) our explanation for the existence of sharecropping in terms of the nonlinearity in supervision cost with respect to land under sharecropping.

Conclusion

In this paper, by introducing the possibility of nonlinearities in supervision costs into a model of contractual choice, we show that systematic variation in the incidence of sharecrop contracts can be explained as rational choice resulting from landowner-worker communication. Hypotheses concerning the effect of supervision costs on contractual mix and the number of supervisors are developed. Empirical tests are conducted, utilizing data from a sample of cotton plantations in Georgia in 1911. The results are consistent with economies of scope and scale in supervision. Furthermore, the tests indicate the existence of nonlinearities in supervision, thus perhaps resolving what has remained a paradox in the literature on contractual choice in agrarian societies, namely, the persistent use of agricultural share contracts.

Appendix

The Model

Let $\bar{N}$ be the endowment of lab
be allocated between wage lab
where n_s and T_s^d are the labor
under share contract, respectiv
to be divided between landlor
ment of both land and labor

$l - l$

In this formulation subscripts refer to the farm type, where w represents the landlord's farm under wage contract and s a sharecropped farm. Superscripts s and d represent supply and demand, respectively, and bars over the variables indicate exogenous variables.

The production functions on landlord and tenant farms are assumed to be identical and of the usual neoclassical type, such that output per unit of land on farm type k, y_k, can be expressed as

$$y_k = f(n_k), \tag{A3}$$

where $k = w, s$.

Let Π = the probability of detection of shirking on the part of the sharecropper, λ = the extent of shirking per unit of contracted labor under wage contract, γ = the penalty of being detected, and c = the supervision cost per unit of land. Suppose also that (1) Π for each type of contract is positively related to both λ (or, in the case of sharecrop contracts, inversely related to n_s) and the c specific to that type of contract, (2) γ is positively related to shirking per unit of land λn_w and negatively related to n_s, and (3) c_s is either positively or negatively (which is more likely to be the case) related to the amount of land T_s^s or T_s^d. More specifically and consistent with the historical evidence on the nature of transactions costs on different types of farms (Alston and Higgs 1982), we assume:

$$\Pi_w = \Pi_w(\lambda, c_w), \tag{A4a}$$

$$\Pi_s = \Pi_s(n_s, c_s), \tag{A4b}$$

where $\partial\Pi_w/\partial\lambda$, $\partial\Pi_w/\partial c_w > 0$, $\partial\Pi_s/\partial n_s < 0$, $\partial\Pi_s/\partial c_s > 0$, $\partial^2\Pi_w/\partial\lambda^2$, $\partial^2\Pi_s/\partial n_s^2 > 0$, and $\partial^2\Pi_w/\partial c_w^2$, $\partial^2\Pi_s/\partial c_s^2 < 0$;

$$\gamma_w = \gamma_w(\lambda n_w), \tag{A5a}$$

$$\gamma_s = \gamma_s(n_s), \tag{A5b}$$

where γ_w', $\gamma_w'' > 0$ and $\gamma_s' < 0$, $\gamma_s'' > 0$;

$$c_w = c_w(n_w), \tag{A6a}$$

$$c_s = c_s(T_s^i), \quad i = s, d \tag{A6b}$$

where c_w', $c_w'' \gtreqless 0$, $c_s' \gtrless 0$, $c_s'' \gtreqless 0$.

Assuming for simplicity that both parties are risk neutral and utilizing the resource constraints (A1) and (A2), the objective functions of the two agents can thus be represented as

$$\begin{aligned}\max_{(T_s^d, n_s, \lambda)} EI_T = {} & (\bar{N} - T_s^d n_s)[(1 + \lambda)\bar{W} - \Pi_w(\lambda, c_w)\gamma_w(\lambda n_w)] \\ & + T_s^d\{\alpha f(n_s) - \Pi_s[n_s, c_s(T_s^d)]\gamma_s(n_s)\};\end{aligned} \tag{A7}$$

$$\begin{aligned}\max_{(T_s^s, n_w)} EI_L = {} & (\bar{T} - T_s^s)[f(n_w - \lambda n_w) - \bar{W} n_w - c_w(n_w)] \\ & + T_s^s[(1 - \alpha)f(n_s) - c_s(T_s^s)],\end{aligned} \tag{A8}$$

where EI_T and EI_L are the expected incomes of the representative tenant and landlord, respectively; α is the tenant's share of output under sharecropping; $\bar{W}$ is the exogenously given wage rate; and $\lambda\bar{W}$ is the income the worker-tenant receives while moonlighting elsewhere as a wage laborer.

The first-order conditions for the tenant are as follows:

$$\begin{aligned}\frac{\partial EI_T}{\partial T_s^d} = {} & -[(1 + \lambda)\bar{W} - \Pi_w\gamma_w]n_s + [\alpha f(n_s) - \Pi_s\gamma_s] \\ & - T_s^d \frac{\partial\Pi_s}{\partial c_s} \cdot c_s'\gamma_s = 0;\end{aligned} \tag{A9a}$$

$$\begin{aligned}\frac{\partial EI_T}{\partial n_s} = {} & -[(1 + \lambda)\bar{W} - \Pi_w\gamma_w]T_s^d \\ & + T_s^d\left[\alpha f'(n_s) - \Pi_s\gamma_s' - \gamma_s \frac{\partial\Pi_s}{\partial n_s}\right] = 0;\end{aligned} \tag{A9b}$$

$$\frac{\partial EI_T}{\partial\lambda} = (\bar{N} - T_s^d n_s)\left(\bar{W} - \Pi_w\gamma_w' n_w - \gamma_w \frac{\partial\Pi_w}{\partial\lambda}\right) = 0. \tag{A9c}$$

Condition (A9a) implies that the net benefits (inclusive of benefits and costs of shirking) per unit of land must be equal for the two types of contracts. Condition (A9b) implies that the marginal net returns (again inclusive of the benefits and costs of shirking) to labor input under the two types of contracts must be equal. Finally, condition (A9c) implies that the marginal benefit from shirking as a wage laborer—that is, the wage rate earned moonlighting—is equal to the marginal cost of such shirking.

The first-order conditions of the landlord are

$$\begin{aligned}\frac{\partial EI_l}{\partial T_s^r} = {} & -[f(n_u - \lambda n_w) - \overline{W}n_u - c_w] \\ & + [(1 - \alpha)f(n_s) - c] - T_s^r c_s' = 0,\end{aligned} \tag{A10a}$$

$$\frac{\partial EI_L}{\partial n_w} = (\overline{T} - T_s^r)\{f'[(1 - \lambda)n_u] - \overline{W} - c_u'\} = 0. \tag{A10b}$$

Condition (A10a) implies that the net earnings per unit of land—after adjustment for the scale effect in supervision of sharecrop contracts—must be equal under the two types of contracts. Likewise, condition (A10b) implies that the net benefits over costs of an extra unit of n_w should be zero. Note that with appropriate and reasonable additional restrictions with respect to the signs and magnitudes of c_s', c_s'', and c_u', the second-order conditions for both parties can be fulfilled

The system is closed by imposing the equilibrium condition

$$T_s^d - T_s^r \tag{A11}$$

Given our assumptions, the system is now fully determined Thus, for given values of the resource constraints and parameters in the model, unique solutions for all eight endogenous variables (T_s^d, T_s^r, n_s, α, T_r, N_u^r, n_u, and λ) and also for γ_u, γ_s, c_u, c_s, Π_s, Π_u, I_f, and I_l can be obtained.

In order to be sure that the model's determinate solution involves no inefficiency, note that conditions (A9a) and (A9b) together imply

$$\alpha[f(n_s) - n_s f'(n_s)]$$

Since the term in brackets on
product of land under the shar
cropping not to be ruled out
positive for sharecropping to exist
are negative, both the first and third te
Note that $c_s' > 0$; that is, diseconomies of scale in supervision, appearing in the second term on the right-hand side, would be sufficient but not necessary to ensure that the marginal product of sharecropped land would be positive

Likewise, equations (A9b) and (A10b) imply the following relationship between n_s and n_w:

$$\alpha f'(n_s) - \Pi_s\gamma_s' - \gamma_s\frac{\partial\Pi_s}{\partial n_s} + \Pi_u\gamma_u = (1 + \lambda)\{(1 - \lambda)f'[(1 - \lambda)n_u] - c_u'\} \tag{A13}$$

Clearly, $n_s \gtreqless n_w$; moreover, with transactions costs, equality between n_s and n_w is not a necessary requirement for efficiency

Three special cases of the formulation above can now be easily investigated. In case A (without supervision on either type of contract), with $\Pi_s = \gamma_s = \Pi_w = \gamma_w = c_s = c_w = 0$, the model reduces to that of Bardhan and Srinivasan (1971) and the right-hand side of equation (A12) becomes zero, thereby implying that the marginal product of land under the share contract, $f(n_s) - n_s f'(n_s)$, must be zero. Likewise, from (A13), it would be seen that $\alpha f'(n_s) = f(n_w)$, implying that $n_s < n_w$ and thus that inefficiency exists. These were indeed the shortcomings of the Bardhan and Srinivasan (1971) model noted by Newbery (1974), Bagchi (1975), and Reid (1976). No interior solution is possible in this case. In case B (with supervision only on wage labor), where $\Pi_s = \gamma_s = c_s = 0$, which may be regarded as a somewhat more competitive and simplified version of the Lucas (1979) model, condition (A12) again reveals the marginal product of land under sharecropping to be zero, a solution inconsistent with the persistence and pervasiveness of sharecropping. In case C (with supervision only on sharecroppers), where $\Pi_w = \gamma_w = c_w = 0$, condition (A12) remains unaffected and hence allows the marginal product of sharecropped land to be positive. But condition (A9c) can be satisfied only with $T_w = \bar{N} - T_s^d n_s = 0$. Once again, no interior solution is possible in this case.

With supervision costs on both wage labor and sharecroppers, therefore, both the Scylla of indeterminacy and the Charybdis of inefficiency can be avoided.

References

Alchian, Armen A., and Demsetz, Harold. "Production, Information Costs, and Economic Organization." *A.E.R.* 62 (December 1972): 777–95.

Alston, Lee J. "Tenure Choice in Southern Agriculture, 1930–1960." *Explorations in Econ. Hist.* 18 (July 1981): 211–32.

Alston, Lee J., and Higgs, Robert. "Contractual Mix in Southern Agriculture since the Civil War: Facts, Hypotheses, and Tests." *J. Econ. Hist.* 42 (June 1982): 327–53.

Bagchi, Amiya K. "Crop-sharing Tenancy in Agriculture: A Rejoinder." *Econ. and Polit. Weekly* 10 (July 26, 1975): 1131–33.

Bardhan, Pranab K., and Srinivasan, T. N. "Cropsharing Tenancy in Agriculture: A Theoretical and Empirical Analysis." *A.E.R.* 61 (March 1971): 48–64.

Ben-Porath, Yoram. "The F-Connection: Families, Friends, and Firms and the Organization of Exchange." *Population and Development Rev.* 6 (March 1980): 1–30.

Braverman, Avishay, and Stiglitz, Joseph E. "Sharecropping and the Interlinking of Agrarian Markets." *A.E.R.* 72 (September 1982): 695–715.

Brooks, Robert Preston. *The Agrarian Revolution in Georgia, 1865–1912.* Madison: Univ. Wisconsin, 1914.

Calvo, Guillermo A., and Wellisz, Stanislaw. "Supervision, Loss of Control, and the Optimum Size of the Firm." *J.P.E.* 86 (October 1978): 943–52.

Cheung, Steven N. S. *The Theory of Share Tenancy.* Chicago: Univ. Chicago Press, 1969.

Hayami, Yujiro, and Kikuchi, Masao. *Asian Village Economy at the Crossroads. An Economic Approach to Institutional Change.* Baltimore: Johns Hopkins Univ. Press, 1982.

Higgs, Robert. "Patterns of Farm Rental in the Georgia Cotton Belt, 1880–1900." *J. Econ. Hist.* 34 (June 1974): 468–82.

Jensen, Michael C., and Meckling, William H. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *J. Financial Econ.* 3 (October 1976): 305–60.

Kotwal, Ashok. "The Impact of the Green Revolution on Labor Contracts and Rural Poverty." Discussion Paper no. 81-33. Vancouver: Univ. British Columbia, Dept. Econ., 1981 (*a*)

———. "Theory of Incentive Contracts in Agriculture—Consumption Credit as a Risk Instrument." Discussion Paper no. 81-31. Vancouver: Univ. British Columbia, Dept. Econ., 1981. (*b*)

Lucas, Robert E. B. "Sharing, Monitoring, and Incentives: Marshallian Misallocation Reassessed." *J.P.E.* 87 (June 1979): 501–21.

Newbery, David M. G. "Cropsharing Tenancy in Agriculture: Comment." *A.E.R.* 64 (December 1974): 1060–66

———. "Tenurial Obstacles to Innovation." *J. Development Studies* 11 (July 1975): 263–77.

Newbery, David M. G., and Stiglitz, Joseph E. "Sharecropping, Risk Sharing and the Importance of Imperfect Information." In *Risk, Uncertainty and Agricultural Development*, edited by James A. Roumasset, Jean-Marc Boussard, and Inderjit Singh. Berkeley. Univ. California Press (for Agricultural Development Council), 1979.

North, Douglass C. *Structure and Change in Economic History*. New York: Norton, 1981.

Pant, Chandrasekhar. "Tenancy and Family Resources: A Model and Some Empirical Analysis." *J. Development Econ.* 12 (February/April 1983): 27–39

Pollak, Robert A. "A Transaction Costs Approach to Families and Households." Mimeographed. Philadelphia: Univ. Pennsylvania, Dept. Econ., 1982.

Reid, Joseph D., Jr. "Sharecropping and Agricultural Uncertainty." *Econ. Development and Cultural Change* 24 (April 1976): 549–76

Roumasset, James A. *Rice and Risk: Decision Making among Low-Income Farmers*. Amsterdam, North-Holland, 1976

Roumasset, James A., and Uy, Marilou. "Piece Rates, Time Rates, and Teams: Explaining Patterns in the
and Organization 1 (December

Stiglitz, Joseph E. "Incentives, R
of Hierarchy." *Bell J. Econ.* 6

Williamson, Oliver E. "Hierarchic
75 (April 1967): 123–38.

———. *Markets and Hierarchies*

———. "The Modern Corporat
Literature 19 (December 1981)

Unionization and Profitability: Evidence from the Capital Market

Richard S. Ruback
Massachusetts Institute of Technology

Martin B. Zimmerman
University of Michigan

This paper examines the effect of unionization on the profitability of firms. Abnormal monthly common stock returns for a sample of 253 NYSE-listed firms are estimated for the month in which the union petitions for an election and for the month in which the National Labor Relations Board certifies the election outcome. The results suggest that unionization, on average, is associated with a reduction in equity value. When unions win an election, the average loss associated with the unionization drive is 3.8 percent of equity value. When unions lose an election, there is an average net reduction of 1.3 percent in the equity value of the firm.

I. Introduction

Unionization can have a variety of effects on the performance of firms. Several of these effects, such as raising wages, introducing inefficient labor practices, and reducing managerial prerogatives, suggest that unionization raises labor costs and thereby reduces profits. For example, Lewis (1981) reviews the literature on wages and

---

We are indebted to Paul Asquith, Fischer Black, Peter Dodd, Henry Farber, Robert Holthausen, Tom Kochan, Richard Leftwich, Susan Martin, Wayne Mikkelson, Walter Oi, Jerry Zimmerman, and participants at workshops at MIT, the University of Chicago, the University of Rochester, the Harvard Business School, and the University of Michigan.

[*Journal of Political Economy*, 1984, vol. 92, no. 6]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9206-0007$01.50

finds that unions raise wages by about 10 percent to 20 percent. Freeman and Medoff (1979), however, suggest that other effects of unionization can increase productivity and lower labor costs. Freeman (1980) finds that unionization reduces quit rates and labor turnover, which increases productivity and lowers the cost of training workers. Additionally, Thompson (1980) argues that unionization limits labor supply, which reduces output and raises product prices and profits.

Prior empirical research has generally focused on a specific effect of unionization and therefore does not measure the net impact of unionization. Some recent research suggests that productivity increases compensate for the higher wages and lead to lower labor costs. Brown and Medoff (1978), using two-digit SIC industry data, relate value added per worker to capital-labor ratios and the proportion of workers unionized. Their results suggest that unionization increases productivity by approximately 20 percent. When compared to the average increase in wages associated with unionization, these empirical estimates support the hypothesis that unions increase profitability.[1] While Brown and Medoff adjust for differences in labor quality and in capital-labor ratios, their estimates are subject to possible errors because their dependent variable, value added per worker, reflects both price and quantity. Therefore, as the authors note, their union effect could be due to higher costs and hence higher product prices in unionized industries. Also, aggregation of essentially different production functions to a single manufacturing production function introduces statistical problems that make their results difficult to interpret.

Clark (1980) circumvents many of these issues by examining performance in cement industry plants. [illegible]
Clark measures output in phy [illegible]
neous technology and avoid [illegible]
sectional investigation. His re [illegible]
productivity, and the net effect [illegible]
and on profits is negligible.[2]

The results of Brown and Me [illegible]
that unionization does not increase [illegible]

---

[1] The Brown and Medoff results suggest an increase in output per worker, with capital-labor ratios held constant, of approximately 20 percent. They find union wages, other things constant, are also approximately 20 percent higher. Labor represents about 70 percent of cost, so that unionization reduces cost per unit. Brown and Medoff do not discuss this aspect of their results.

[2] In an unpublished study of the construction industry, Allen (1982) finds no credible effect of unions on productivity in school construction. In office construction Allen finds productivity is 38 percent higher for union labor. He finds a roughly equal effect on wages, which implies higher profits. Output measurement problems are more severe here than in Clark (1980).

are intriguing since they suggest that unionization does not reduce and may increase profits; union-induced increases in productivity seem to offset higher union wages. However, casual evidence indicates that management typically resists unionization. Since managerial resistance to unionization is costly, the result that unions do not increase production costs implies that the resistance is inconsistent with stockholder wealth maximization.

This paper directly examines the effect of unionization on profitability. Instead of extracting the effect of unionization through productivity and wage data, the unanticipated change in the equity values of firms involved in union elections is used as a measure of the effect of unionization on the capitalized value of the firms' profits.

This approach has several advantages. First, there is ample evidence that an unbiased assessment of the effects of public information releases about firm profitability is quickly incorporated into stock prices.[3] This eliminates the necessity of specifying production functions and the length of time required to adjust factor inputs. Higher wages, not compensated by increases in productivity, reduce the present value of profits and the equity value of the firm falls. If unionization increases productivity to offset higher wages, profits increase and equity values rise. Second, since profits include the effects of both higher wages and higher product prices, the net effect of unionization can be measured without ambiguity. If unions succeed in limiting entry to the industry, this is also reflected in the revaluation of the firm's equity. The method allows aggregation across firms and industries, which expands the sample size and permits analysis that is not industry specific.

The approach has several disadvantages stemming from the efficiency of the stock market. First, to the extent that unionization is anticipated, the measured unanticipated change in equity values associated with a representation election will underestimate the actual impact of unionization. Second, the outcome of a representation election may provide information about the outcome of future representation elections. Thus the loss or gain caused by a single election may reflect more than the impact of the union on the productivity and wages of the newly organized workers. Also, the estimation technique focuses on new representation elections and measures the marginal impact of unionization on profits. It is inappropriate, therefore, to interpret the results as measuring the aggregate impact of unionization on corporate profits.

The empirical results indicate that on average unionization is associated with a reduction in profitability. Stockholders of firms that are unionized incur losses of about 4 percent of equity value. In firms whose workers petition for a union election but decide not to become unionized, equity values decline by about 1 percent.

---

[3] See Fama (1976) for a survey of market efficiency tests.

II. The Unionization Process

There are three important steps in the union organization process. The first official step in the process is the *petition* for an election. To file a petition, the union must get at least 30 percent of the workers at the bargaining unit to sign cards authorizing the union to represent them. Once 30 percent of the workers have signed, the union petitions the National Labor Relations Board (NLRB) to hold an election. If more than 50 percent of the workers sign cards, the union can ask the firm directly for recognition. Typically, recognition is denied and the union petitions the NLRB for an election.

The petition confirms that a union representation election will occur and thus raises the probability of unionization. In an efficient capital market, equity prices will be revised to reflect the probability of unionization and the expected effects of the unionization on the present value of future cash flows. If unionization decreases profits, petitions will be associated with a decline in the stock price of the firm. Alternatively, if unionization increases profitability, stock prices will rise in response to the petition.

The second step in the process is the representation election itself. In most cases, where union and management agree on what constitutes the bargaining unit, the election is held shortly after petition. Roomkin and Block (1981) find that 12 percent of elections occurred within 1 month and 83 percent with[illegible] cation of the election results[4] [illegible] days after the election, unless [illegible] from the union or from mana[illegible] Roomkin and Block found pos[illegible] the cases they analyzed.

The certification of the el[illegible] about the outcome of the un[illegible] certified, the probability of unio[illegible] market reaction to the certification[illegible] information about the effect of unio[illegible] unionization decreases profitability, certification of the union's success should cause a reduction in stock prices. Conversely, if unionization increases profitability, the union's success should be associated with an increase in stock prices.

---

[4] Certification of a union win does not resolve all the uncertainty associated with unionization. For example, a contract has to be negotiated, and its terms are, of course, uncertain at the time of certification

If the NLRB certifies that the union has lost the election, the predicted stock price reaction under each hypothesis is reversed. However, there are two reasons why the stock price changes associated with union losses and union wins do not have to be symmetric. First, unions lose most representation elections, and therefore the elections that unions win are likely to involve a larger revision in the markets' ex ante probability of unionization. In other words, a union win is likely to contain more new information than a union loss, so that the magnitude of the abnormal return for union losses will be greater than the stock price revisions associated with union wins.[5] Second, the loss of an election does not prohibit the union from petitioning for another election in the future, and thus the probability of eventual union representation does not become zero.

In summary, the hypothesis that unions decrease profitability implies that the petition for a union election is associated with a decrease in stock prices. The hypothesis also predicts that certification of a union success is associated with a further decline in stock prices, whereas certification of a union loss results in an increase in stock prices. The predicted combined effect of petition and certification should be negative for union success. The hypothesis that unions increase profits predicts exactly the opposite effect.

For both hypotheses, the predicted combined abnormal return on the petition and certification months for an unsuccessful union election depends on the cost of opposing the union and the change in the probability of eventual unionization. The cost of opposition includes direct costs such as legal fees and indirect costs such as higher wages or other concessions granted to workers to make unionization less attractive. The probability of eventual unionization could either increase or decrease as a result of a successful petition and an unsuccessful election. The successful petition indicates some support for

---

[5] Suppose, e.g., that the market assigned the same probability of a union win to each election at the time of petition. If the market is efficient, the average abnormal return across union wins and losses at the certification date should be zero. Defining f_w as the proportion of union wins, R_w as the abnormal return associated with a union win, and R_l as the abnormal return associated with a union loss at certification, $0 = f_w R_w + (1 - f_w) R_l$ and thus

$$\frac{R_w}{R_l} = \frac{-(1 - f_w)}{f_w}.$$

Unions win 21 percent of the elections in our sample. This implies that the abnormal return to union wins at the certification date should be 3.75 times greater than the abnormal return for union losses in absolute value. The details of this example are not, however, consistent with our data since the market appears partially to anticipate the outcome of union elections.

unionization whereas the unsuccessful election indicates that a majority of workers opposed unionization. Given the cost of opposition and the ambiguous combined effect of petition and certification on the probability of eventual unionization, the predicted combined abnormal returns under both hypotheses are ambiguous for unsuccessful unionization attempts.

III. Data

The sample of union elections used in this study is from the files of the NLRB. The sample is limited to unions that have successfully petitioned for a representation election covering at least 750 employees between February 1962 and September 1980. Also, to use the monthly stock return data provided by the Center for Research in Security Prices (CRSP) of the University of Chicago, firms not listed on the New York Stock Exchange (NYSE) are excluded.[6] These sample selection criteria identify large representation elections in large firms.[7] Since our sample excludes other organization drives as well as firms that unions chose not to attempt to organize, our empirical results cannot be applied to any randomly selected firm.

Information about the elections is published in the *NLRB Monthly Election Report*. The report provides data on the number of workers involved in the election, the outcome of the election, and the certification date. The petition date for each election was obtained through a Freedom of Information Request to each district office of the NLRB. The sample is restricted to elections for which petition and certification dates are available.

A given firm can have several separate unionization attempts. In
some cases, these separate atte
organization attempt can occ
date for another organization
with overlapping events, if the
organization drive, the enti
tempts is treated as a single
tion becomes the petition date
certification date. In cases where ele
vations are dropped.

In the period 1962–80 there were 325 elections involving more than 750 workers for NYSE-listed firms in the period. Lack of stock

---

[6] We were able to identify the month of petition and certification for the entire sample. The actual day of the event was available only for a very limited sample.

[7] The National Labor Relations Board (1980) reports that there are about 8,000 representation elections a year. The vast majority of these elections involve fewer than 100 workers.

TABLE 1

Time Profile of Petitions for Union Elections

Year of Petition	Total	Unsuccessful Union	Successful Union	Union Success Ratio (%)*
1961	2	1	1	50.0
1962	7	6	1	14.3
1963	14	11	3	21.4
1964	14	9	5	35.7
1965	15	11	4	26.7
1966	12	10	2	16.7
1967	18	11	7	38.9
1968	20	16	4	20.0
1969	14	12	2	14.3
1970	14	9	5	35.7
1971	15	13	2	13.3
1972	11	9	2	18.2
1973	27	22	5	18.5
1974	22	18	4	18.2
1975	9	9	0	0
1976	5	5	0	0
1977	5	3	2	40.0
1978	12	10	2	16.7
1979	12	10	2	16.7
1980	5	4	1	20.0

* The union success ratio is the number of elections in which the union is successful divided by the total number of elections in the year

market data eliminated 38 observations. The combining of overlapping elections eliminated an additional 34 elections. Thus, the sample consists of 253 elections. Unions lost 199 and won 54 of these elections. Petitions by 40 different unions are contained in the sample. The United Auto Workers appears most frequently in the sample with 29 unsuccessful and eight successful unionization drives. Table 1 presents a time profile of the data. No trend either in the total number of elections or in the union success ratio is apparent in the sample.[8]

IV. Methodology

The event-study methodology pioneered by Fama et al. (1969) is used to measure the effect of petitions and certification of election results on the stock prices of firms.[9] Since most stocks tend to move up or down with the market, the realized stock returns are adjusted for marketwide movements to isolate the component of the returns due to union activity. This adjustment is accomplished using linear regression to estimate the following market model:[10]

---

[8] For comparison, the success ratio for unions in all representation elections has declined from an average of about 59 percent in the early 1960s to an average of about 46 percent in the latter part of the 1970s (National Labor Relations Board 1960, 1980).

[9] This methodology has been used to examine the equity value changes associated with a variety of events including stock splits, dividends, changes in corporate charters, mergers, and tender offers. Neumann (1980) uses this technique to examine the impact of labor strikes on equity values Schwert (1981) reviews applications of this technique.

$$R_t = \alpha_j + \beta_j R_{mt} + \epsilon_t. \qquad (1)$$

The parameter β_j measures the sensitivity of the jth firm's return (R_{jt}) to movements in the market index (R_{mt}) and is equal to the beta of the Sharpe-Lintner capital asset pricing model (CAPM). The term $\beta_j R_{mt}$ in equation (1) is the portion of the return to security j in month t that is due to marketwide factors. The parameter α measures that part of the average return of the stock that is not due to market movements. Last, ϵ_t measures that part of the return to the firm that is not due to movements in the market or the firm's average return.

Two sets of coefficients are estimated for each firm to incorporate potential risk shifts associated with unionization activity. Coefficients before the petition, α^B and β^B, are estimated using 60 months of data beginning 7 years before the petition. Similarly, coefficients after the election are estimated over the period from 25 months after the certification through 84 months after the election. In those cases in which 24 months of data are not available to estimate either the before or after coefficients, combined data from before the petition and after certification are used to estimate the coefficient. In all cases, data for the 24 months preceding the petition through 24 months following the certification are excluded from the estimation period. All of the results presented below were replicated using coefficients calculated from data combining 60 months before the petition and 60 months after the election. The conclusions of this study are not affected by using the pooled coef[illegible]

Prediction errors are calcu[illegible]
months before petition thro[illegible]
the following equation:

$$PE_{jt} = \begin{cases} R_t - (\hat{\alpha}_j^B \text{[illegible]} \\ R_t - (\hat{\alpha}_j^A \text{[illegible]} \end{cases}$$

The term in parentheses in equation (2 [illegible]
for security j in month t, given its estimated market model parameters $\hat{\alpha}$ and $\hat{\beta}_j$ and the market return in month t. The prediction error for firm j in month t is its actual return (R_{jt}) less its predicted return ($\hat{\alpha}$ + $\hat{\beta}_j R_{mt}$) in month t and thus proxies for firm-specific abnormal returns.

In some cases, calculating prediction errors for the 24 months before petition and after certification causes an overlap with another unionization attempt for the same firm. To avoid double counting, the month in which the first prediction error is calculated is the minimum of 24 months prior to petition or the month after the preceding certification. Similarly, the month of the last prediction error is the minimum of 24 months after the certification date or 24 months before the next petition for the firm. If any of the 24 months before the next petition overlaps the certification date, the last date for the calculation of prediction errors is the certification month. This procedure ensures that the prediction error for a given firm in a given month is used at most once in the analysis.

---

[10] Fama (1976) describes the market model in detail

The prediction errors are averaged across observations for months relative to the petition and certification dates:

$$APE_\tau = \frac{1}{J_\tau}\sum_{j=1}^{J_\tau} PE_{j\tau} \quad \tau = PD - 12, \ldots, \\ PD - 1, PD, CD, CD + 1, \ldots, CD + 12, \tag{3}$$

where J_τ is the number of prediction errors in each event month τ, PD is the month of petition, and CD is the month of certification. The time between petition and certification varies across observations. The APE for this interval is calculated as

$$APE_I = \frac{1}{J_I}\sum_{j=1}^{J_I}\sum_{t=PD+1}^{CD-1} PE_{jt}. \tag{4}$$

The average prediction errors represent the average deviation of the monthly returns from their normal relation with the market and proxy for abnormal returns. The average prediction errors are summed over event time to form a cumulative average prediction error (CAPE), which measures the abnormal performance over an interval of event-related time.

To test the statistical significance of the abnormal returns, we compute the following t-statistic:

$$t = \frac{\sum_{j=1}^{J}\left[\sum_{t=\tau_1}^{\tau_2} PE_{jt} \Big/ \sqrt{\sum_{t=\tau_1}^{\tau_2} \mathrm{var}(PE_{jt})}\right]}{\sqrt{J}}, \tag{5}$$

where τ and τ_2 are the first and last months of the cumulation period, J is the number of observations, and $\mathrm{var}(PE_{jt})$ is the variance of the prediction error for firm j in month t. The variance of the prediction errors is calculated as:

$$\mathrm{var}(PE_{jt}) = S_j^2\left[1 + \frac{1}{N} + \frac{(R_{mt} - \bar{R}_{mt})^2}{(N - 1)\mathrm{var}(R_m)}\right], \tag{6}$$

where S_j^2 is the residual variance from the market model regression and N is the number of observations used to estimate the market model.[11] The t-statistic adjusts for heteroscedasticity in the prediction errors by standardizing the cumulative prediction errors for each firm by their standard deviation. This standardization gives less weight to the prediction errors of more volatile firms because the prediction error for such firms conveys less information than does the prediction error of less volatile firms. It is possible for the abnormal returns (as measured by the cumulative average prediction errors) and the t-statistics to be of opposite sign since the standardizing process results in the assignment of nonuniform weights to the individual cumulative prediction errors. This is most likely to occur when the average cumulative prediction errors are close to zero.[12]

V. Results

A. *Petition Date*

Table 2 summarizes the cumulative abnormal returns for selected holding periods for the entire sample and the two subsamples. Also shown in table 2, in parentheses, is the t-statistic for the cumulative abnormal return and, in brackets, the percentage of the observations that have positive cumulative abnormal returns over the holding period.

According to the hypothesis that unionization increases profits, the abnormal return on the petition date should be positive. The hypothesis that unionization decreases profits implies a negative abnormal return. Across all observations, the abnormal return in the petition month, *PD*, is -1.38 perc [illegible]
statistically significant at the [illegible]
significant negative abnormal [illegible]
with the hypothesis that unio [illegible]

The abnormal return in the [illegible]
the subset of unsuccessful u [illegible]
successful union elections wit [illegible]
and -2.93, respectively. The estima [illegible]
unions is twice as large as for unsuccessful unions, [illegible]
of equal impact cannot be rejected at the 95 percent confidence level

---

[11] The variance of PE_{jt} is the usual variance of the prediction error. It is assumed that prediction errors across firms are independent. We calculate the variance of the cumulative prediction errors over event time as the sum of the individual variances. This is only an approximation since it ignores the covariances between prediction errors.

[12] We calculated t-statistics without the heteroscedasticity correction, and the qualitative results are not affected.

TABLE 2

Percentage Cumulative Monthly Average Prediction Errors for Selected Holding Period (t-Statistics in Parentheses and Percentage Positive in Brackets)

Holding Period	*All Observations*	*Unsuccessful Unions*	*Successful Unions*
PD and *CD**	-1.86 (-2.58) [43]	-1.32 (-1.64) [45]	-3.84 (-2.45) [35]
PD - 12 to *PD* - 1	-1.01 (-.43) [47]	.70 (.63) [49]	-7.17 (-2.12) [39]
PD	-1.38 (-3.69) [45]	-1.10 (-2.64) [47]	-2.41 (-2.93) [35]
PD + 1 to *CD* - 1†	.56 (.11) [48]	.56 (-.02) [47]	.53 (.27) [50]
CD	-.48 (-.49) [47]	.22 (.06) [49]	-1.43 (-.95) [39]
PD to *CD*	-1.30 (-1.91) [45]	-0.76 (-1.30) [45]	-3.30 (-1.65) [43]
CD + 1 to *CD* + 12	.34 (.26) [47]	.29 (.26) [46]	.41 (.05) [49]
Number of Observations‡	253	199	54

* *PD* is the month in which the petition is filed and *CD* is the month in which the election results are certified.

† *Petition and certification occur in successive months for many observations. To preserve additivity, we report* the cumulative average return over the period *PD* + 1 to *CD* - 1 as if this interim performance is zero for observations in which petition and certification occur in successive months. The *t*-statistics use the actual average *prediction errors in this period along with the associated sample size as reported in table 3.*

‡ The number of observations varies across holding periods. The maximum number of observations occurs on both *PD* and *CD*. Table 3 contains the number of observations in each event month.

because of the relatively high standard errors.[13] While not statistically significant, the difference in magnitude of the loss for successful and unsuccessful unionization attempts suggests that the market anticipates the outcome of the election. This anticipation is consistent with

---

[13] A *t*-statistic for the difference between the petition month abnormal returns for successful and unsuccessful union elections is calculated by estimating the regression equation:

$$PE_j(PD) = \underset{(0.42)}{-1.17} + \underset{(0.90)}{-1.29}D_j, \quad R^2 = 0.02,$$

where $PE_j(PD)$ is the percentage prediction error for observation j in the petition month, and D_{ij} = 1 for union wins and zero otherwise. The standard errors of the estimated coefficients are in parentheses. The *t*-statistic on the binary variable, -1.44, indicates that the differences between abnormal returns for the samples of union wins and losses are not statistically significant at the 95 percent level of confidence. The regression is estimated using weighted least squares where the weights equal the inverse of the standard deviation of the prediction errors, which is calculated using eq. (6).

the findings of Getman, Goldberg, and Herman (1976) that the percentage of workers signing authorization cards (which is available on the petition date) is a good predictor of the election outcome. Since it appears there is anticipation at the petition date, it is possible that differences in abnormal returns prior to the petition for the successful and unsuccessful union subsamples also reflect anticipations about unionization. The stockholders of firms in which the unions are successful incur losses of 7.17 percent with a t-statistic of -2.12 over the year prior to petition. In contrast, over the same period the abnormal returns for the unsuccessful union subsample is 0.70 percent with a t-statistic of 0.63.

The prior losses for firms in which the unions are successful and the lack of significant abnormal returns for firms in which unions are unsuccessful suggest that the outcome of the unionization attempt is anticipated prior to the petition date. However, the pattern of returns sheds doubt on this interpretation. Table 3 presents the abnormal returns in each month from $PD - 12$ to $CD + 12$ for the entire sample and the successful and unsuccessful subsample. The corresponding cumulative abnormal returns are plotted in figure 1. Table 3 and figure 1b indicate that the prepetition losses for the firms in which unions are successful are concentrated in the period $PD - 12$ through $PD - 7$. The abnormal return for these firms in the 6 months immediately preceding the petition is 0.16 percent. This timing suggests that the prepetition abnormal returns are not due to unionization. Instead, the results suggest that firms in which unions are successful experienced declines in value prior to the union activity.

B. Certification Date

The hypothesis that unions [illegible] certification of successful [illegible] abnormal returns, whereas [illegible] with the certification of [illegible] pothesis that unions increase [illegible] the abnormal returns in the certification [illegible] ence between petition date abnormal returns for successful and unsuccessful unionization attempts suggests that the market anticipates the election outcome. If the outcome is largely anticipated, the abnormal returns, which are unanticipated changes in stock prices, will be relatively small and insignificant.

Union losses are associated with a small and insignificant negative return of -0.22 percent in the certification month. While the return for firms in which the union is successful is relatively large in the certification month, -1.13 percent, the standard error is high and

TABLE 3

MONTHLY AVERAGE PREDICTION ERRORS (*APE*) AND POSITIVE INDIVIDUAL PREDICTION ERRORS (*%POS*) FOR ENTIRE SAMPLE AND SUBSAMPLES OF SUCCESSFUL AND UNSUCCESSFUL UNION ELECTIONS (12 Months before Petition [*PD*]–12 Months after Certification)

PERIOD	ENTIRE SAMPLE			UNSUCCESSFUL UNIONS			SUCCESSFUL UNIONS		
	APE (%)	*N*	*%POS*	*APE* (%)	*N*	*%POS*	*APE* (%)	*N*	*%POS*
PD − 12	−.80	211	45	−.66	166	48	−1.30	45	37
PD − 11	−.91	214	45	−.79	169	47	−1.36	45	37
PD − 10	−.40	221	44	.08	174	50	−2.21	47	25
PD − 9	−.66	221	41	−.75	173	39	−.35	48	47
PD − 8	−.70	224	47	−.39	175	48	−1.82	49	46
PD − 7	−.21	225	48	.11	176	50	−1.36	49	40
PD − 6	.32	228	51	.31	178	50	.38	50	56
PD − 5	−.11	231	50	.10	179	52	−.86	52	44
PD − 4	.69	234	50	.95	182	52	−.18	52	44
PD − 3	1.08	238	49	1.19	185	50	.68	53	45
PD − 2	.02	242	50	−.02	188	48	1.00	54	57
PD − 1	.14	249	47	.41	195	48	−.85	54	42
PD	−1.38	253	44	−1.10	199	47	−2.41	54	35
PD + 1 to *CD* − 1	.98	144	47	1.00	112	47	.90	32	50
CD	−.48	253	47	−.22	199	49	−1.43	54	38
CD + 1	.47	187	47	.66	140	49	−.10	47	42
CD + 2	−1.03	184	42	−1.09	138	40	−.85	46	47
CD + 3	−.45	184	44	−.79	138	42	.57	46	47
CD + 4	−.73	184	42	−.37	138	41	−1.80	46	45
CD + 5	−.11	183	45	−.70	137	43	1.63	46	54
CD + 6	1.31	181	54	1.36	135	55	1.14	46	50
CD + 7	−.66	180	43	−.09	134	40	.03	46	52
CD + 8	.36	177	46	.65	132	48	−.48	45	42
CD + 9	.70	177	47	1.00	132	50	−.18	45	39
CD + 10	.45	175	47	.62	130	49	−.05	45	42
CD + 11	−.46	173	45	−.26	128	45	−1.00	45	44
CD + 12	.56	172	51	.20	127	48	1.56	45	57

the loss is not statistically significant (t-statistic of -0.95). These insignificant abnormal returns are consistent with the earlier results indicating that the outcome of the election is largely anticipated at the petition date.[14]

The abnormal returns in the certification month support the efficient market hypothesis. This hypothesis maintains that, at the time of the petition, stock prices incorporate the probabilities of successful and unsuccessful union elections and their concomitant losses. This implies that a trading rule based on information available at the time of the petition, such as purchasing the equities of firms in which unions petition for an election, will not, on average, generate abnormal returns. The results presented in table 2 are consistent with the efficient markets hypothesis, since the average abnormal return for the entire sample at certification is only 0.48 percent. The lack of substantial abnormal returns at certification suggests that the market formed unbiased estimates of the probability of union success and the effects of unionization on future net cash flows at the time of petition.

C. *The Interim Period and Cumulative Loss*

The abnormal returns over the period *PD* through *CD* are -1.30 percent for the whole sample, -0.76 percent for firms involved in unsuccessful unionization attempts, and -3.30 percent for union successes. While consistent with the hypothesis that unionization decreases profits, these results are not statistically significant. The lack of significance seems to be related to the period between petition and certification. As shown in table 2, the abnormal returns in this period are approximately 0.5 percent for both union successes and failures. The abnormal returns during

subsets.

The performance in the p

based on a small sample of f

successive months for most

there is some ambiguity in

---

[14] However, 61 percent of the firms where

mal returns in the certification month, which is significantly different from

at the 90 percent confidence level using a nonparametric sign test

$$= \frac{N(P - H)}{\sqrt{NH(1 - H)}}$$

Here N is the number of observations (54), H is the proportion of negative observations under the null hypothesis (0.50), and P is the proportion of negative observations that occur in the sample (0.61). The resulting t-statistic is 1.62. This is weak evidence, along with the relatively high negative average abnormal return for union wins, that the outcome is not completely anticipated at the petition date.

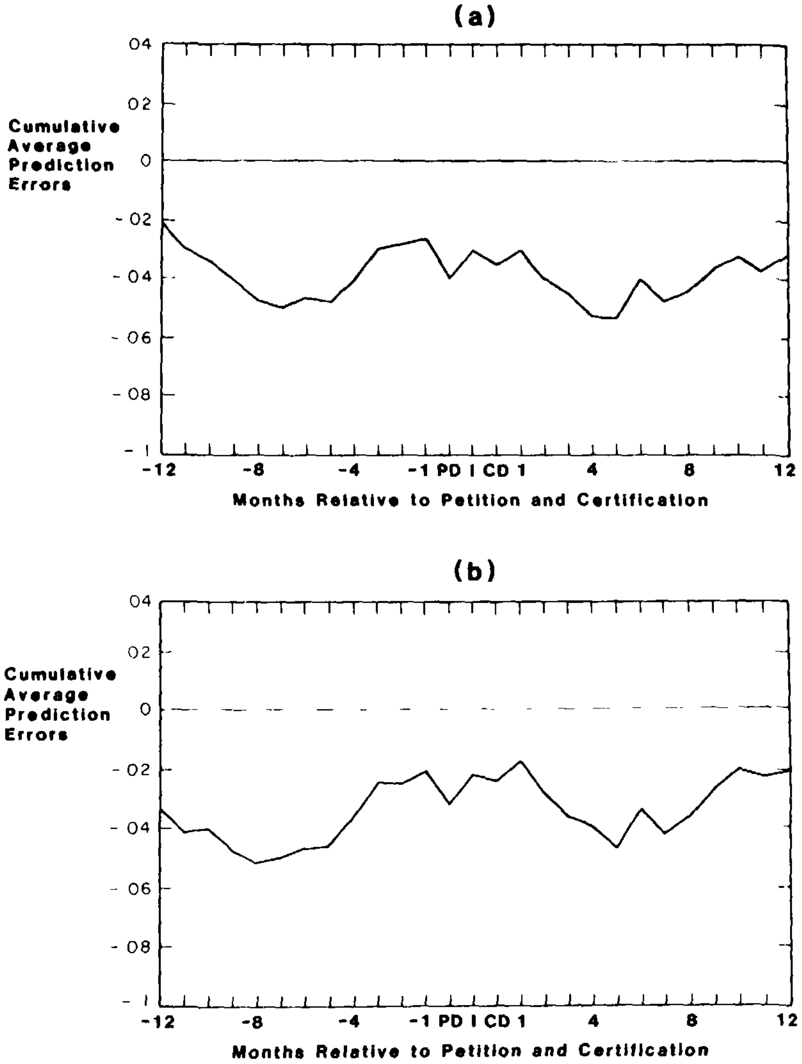


FIG. 1.—Plot of cumulative average prediction errors over 12 months before the petition month (*PD*) through 12 months after the certification month (*CD*). *a*, Whole sample *b*, 199 unsuccessful union elections. *c*, 54 successful union elections.

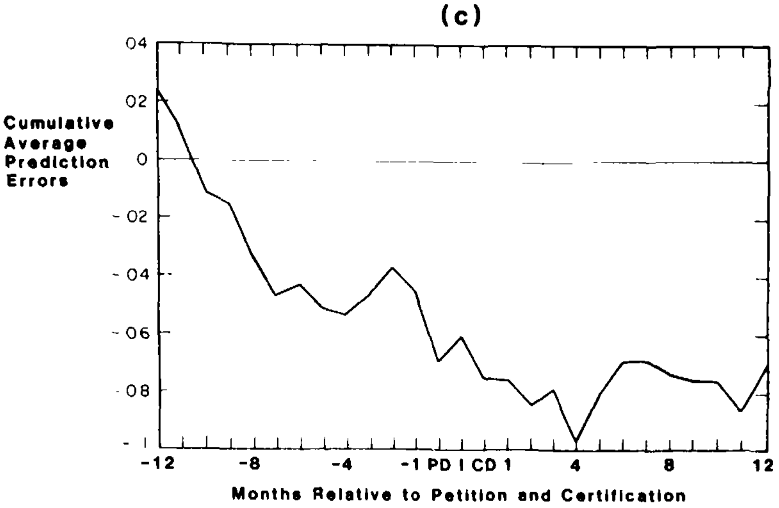


Fig. 1.—*Continued*

period. In some cases the intermediate period includes petitions, elections, and certification months for other ongoing unionization attempts. Because of the overlapping organizational drives, we focus on the first petition and the last certification. Apparently, including the period between petition and certification increases the variance substantially but provides little information.

The difficulties associated with the period between petition and certification suggest that a better me [illegible] effect of union activity is the combined effe
dates. For the entire sampl
activity is a loss to the sha
percent with a *t*-statistic of
the combined effect of petiti
a *t*-statistic of −1.64. The
−3.84 percent with a *t*-statist

The significant negative combin
tion and certification dates for the entire sample and for the sample of successful unionization attempts are consistent with the hypothesis that unionization reduces profitability. Furthermore, the losses for firms in which unions are successful are about three times larger than the losses with unsuccessful unions. While the standard errors of the estimated abnormal returns are too large to reject the hypothesis of equal impact, the relative magnitude of the abnormal returns for successful and unsuccessful unions corroborates the

significant losses for the entire sample on the petition date and the significant losses for successful unions.[15] The evidence therefore is entirely consistent with the hypothesis that unionization is associated with lower profits. The following section explores the differential impacts of the successful and unsuccessful unionization attempts in more detail.

An alternative way to present the results is in terms of the average dollar loss per worker involved. The dollar loss per worker for each firm is calculated by multiplying the abnormal returns associated with union activity (the sum of the *PD* and *CD* prediction errors) times the equity value in the month prior to petition and dividing by the number of workers in the bargaining unit. These losses are converted to 1980 dollars (using the CPI) and averaged across observations. The average loss per worker standardizes our results for the size of the firm and number of workers in the bargaining unit. The average loss per worker therefore measures the present value of a unionized worker's effect on profits. The average loss per worker is −$15,820 for the entire sample. For unsuccessful unions the average loss per worker is −$7,410, and the average loss per worker is −$46,810 for successful unionization attempts. The difference between the average loss per worker in the successful and unsuccessful union subsamples is consistent with the hypothesis that unionization decreases profitability.[16]

Simple calculations suggest that the estimates of loss per worker for successful union elections are at least plausible. In 1980, for example, the average difference between union and nonunion manufacturing wages was $2,184 per year. If this represents a perpetual payment per

---

[15] A *t*-statistic for the difference between the combined petition and certification month abnormal returns for the sample of successful and unsuccessful union elections is calculated by estimating the following regression equation

$$PE_j(PD + CD) = \underset{(0.63)}{-0.91} - \underset{(1.40)}{1.91}D_j, \quad R^2 = 0.02,$$

where $PE_j(PD + CD)$ is the sum of the percentage prediction errors for firm *j*, and $D_j = 1$ for union wins and zero otherwise. The standard errors of the estimated coefficients are in parentheses. The *t*-statistic on the binary variable, −1.36, indicates that the difference between the abnormal returns for the sample of union wins and losses is not statistically significant. The regression is estimated using weighted least squares where the weights equal the inverse of the standard deviation of the prediction errors, which is calculated using eq. (6).

[16] The *t*-statistics for the loss per worker are the same as the *t*-statistics presented for the abnormal returns in the petition and certification months (first line of table 3). The loss per worker for a given firm is the abnormal return on the petition and certification months times the equity value per worker. The standard deviation of the loss per worker is the standard deviation of the abnormal returns times equity value per worker. Since equity value per worker appears in both the numerator and denominator, it cancels in the standardization process.

worker, at a 5 percent real discount rate, the present value of incremental wages is $43,680. This calculation is, of course, only approximate because the appropriate wage differential should be adjusted to reflect the quality of workers. Also, the appropriate interest rate to use in discounting the wage differential is difficult to determine. Plausible discount rates range from 2 percent to 10 percent, which yields a range of loss per worker of about $22,000 to $109,000. It is possible that the measured loss per worker associated with a successful union election is due in part to increased expectations of future unionization attempts in unorganized bargaining units within the firm. The correct measure of loss per worker would then be lower than measured here. However, there is a possible bias in the opposite direction since the measured loss per worker assumes that unionization was not anticipated prior to the petition month.

The pattern of abnormal returns is consistent with the view that unions cause losses. There are, however, other explanations consistent with the results. For example, one interpretation is that stock prices fall when the union petitions for an election because it signals that the firm has poor labor relations and poor labor relations lead to decreased profitability. Similarly, the revision in equity values when the election results are certified is due to additional information about the firm's labor relations. An unsuccessful unionization attempt implies that the firm's labor relations are better than anticipated or that the problems have been resolved. A successful union election indicates that the firm's labor relations are worse than previously anticipated.

While this hypothesis is plausible and consistent with the results, it
is an unlikely explanation for the [illegible] re-
turns. If labor relations are [illegible]
ways. For example, high rat[illegible]
tivity, and lower earnings w[illegible]
union election would reveal [illegible]
the labor relations and its [illegible]

D. *Specific Industry and Union Effects*

The abnormal returns presented in table 3 indicate that unionization is associated with a significant decline in equity values. There is, however, wide dispersion in the abnormal returns across observations. Figure 2 contains a histogram of the combined abnormal returns on the petition and certification dates for the entire sample and for the subsamples of union wins and losses. While the distributions appear to be symmetric, the range of the abnormal returns is substantial. The sum of the prediction errors on the petition and certification dates for

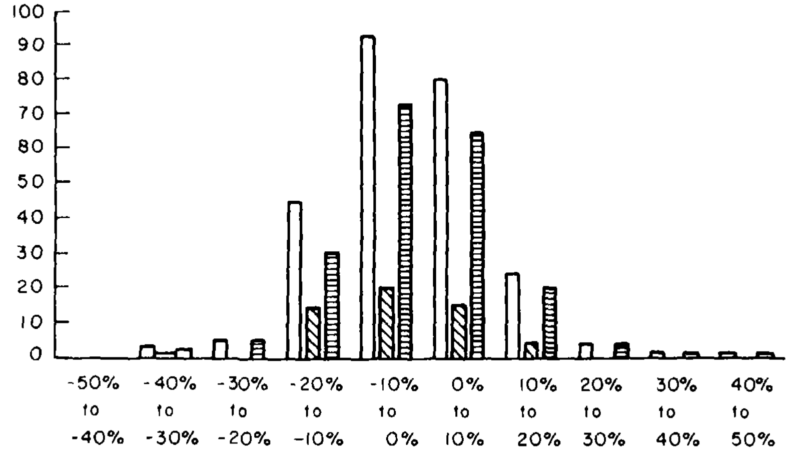


FIG. 2.—Histogram of combined abnormal returns on petition and certification date for the whole sample and subsample of union wins and union losses. Within each cell the first bar (blank) relates to the whole sample, the second bar (sloping lines) relates to union wins, and the third bar (horizontal lines) relates to union losses.

union wins ranges from −36 percent to 17 percent and the abnormal returns for union losses range from −39 percent to 42 percent. One plausible explanation for this wide dispersion of the abnormal returns is that the effects of unionization differ systematically across unions and industries.

To test for particular union and industry effects, abnormal returns are regressed on a series of binary variables representing different unions and different industries.[17] Table 4 shows the results with separate binary variables included for wins and losses for the five unions most frequently appearing in the sample: the United Auto Workers, the Teamsters, the United Steelworkers, the International Union of Electrical, Radio, and Machine Workers, and the International Association of Machinists and Aerospace Workers. These unions represent over 25 percent of all organized workers in the United States (Gifford 1982). In addition, a separate binary variable with a value of one for each union win is included. Since the data seem to indicate that the outcome of the election is partially anticipated at the petition date, two regressions are estimated. The dependent variable in the first regression is the percentage prediction error on the petition date. The dependent variable in the second regression is the sum of percentage prediction errors on the petition and certification dates. To

---

[17] Separate binary variables are included for each union and industry that have at least 20 observations in the sample.

TABLE 4

EFFECTS OF UNION WINS AND LOSSES FOR VARIOUS UNIONS ON THE PETITION AND COMBINED PETITION AND CERTIFICATION DATE ABNORMAL RETURNS (*t*-Statistics in Parentheses)

$$Y = a_0 + \sum_{k=1}^{11} a_k D_{jk} + \epsilon_j$$

INDEPENDENT VARIABLES	Petition Date Abnormal Returns (%)	Combined Petition and Certification Date Abnormal Returns*	*N* OBSERVATIONS
	DEPENDENT VARIABLES		
Constant	-1.12 (2.09)	-1.07 (-1.22)	253
Union win	1.29 (-1.14)	-.84 (-.46)	54
United Auto Workers loss	1.26 (.94)	.72 (.36)	29
United Auto Workers win	-1.80 (-.72)	-6.11 (-1.65)	8
Teamsters loss	1.00 (.59)	3.05 (1.49)	16
Teamsters win	2.91 (.79)	2.00 (.38)	4
International Electrical, Radio, and Machine loss	-.92 (-.65)	-.51 (-.25)	18
International Electrical, Radio, and Machine win	-6.32 (-1.41)	-8.99 (1.47)	1
United Steel Workers of America loss			
United Steel Workers of America win			
Machinist and Aerospace loss			
Machinist and Aerospace win			
R^2			

NOTE.—The five unions that appear most freq [illegible] defined as follows D_1 = 1 for union wins and zero otherwise [illegible] for United Auto Workers and zero otherwise D = 1 for a successful election [illegible] otherwise, D_1 = 1 for a Teamsters loss and zero otherwise D = 1 for a Teamsters win and zero otherwise [illegible] on. The regressions are estimated using weighted least squares where the weights are the inverse of the standard deviation of the dependent variable

* The abnormal return is measured as the sum of prediction errors in the petition and certification months

correct for heteroscedasticity, the regressions are estimated using weighted least squares where the weights are the inverse of the standard deviation of the abnormal returns for each individual firm.

The estimated intercepts in table 4 measure the average abnormal return associated with unsuccessful unionization attempts by unions other than the five most frequently appearing in the sample. The estimated intercept is -1.2 percent with a t-statistic of -2.09 for the petition month abnormal return regression and -1.1 percent with a t-statistic of -1.22 for the combined petition and certification month abnormal return regression. The union win coefficient measures the difference between successful and unsuccessful election for unions other than those examined individually. This coefficient is negative and insignificant in both regressions.

The coefficients for the five unions appearing most frequently in the sample vary widely, suggesting the presence of specific union effects. The large standard errors of these coefficients, however, prevent strong inferences. For example, the coefficient for an International Union of Electrical, Radio, and Machine Workers win is -9.0 percent in the combined petition and certification month abnormal return regression. This coefficient indicates that firms in which this union is successful incur a 9.0 percent decline in equity values beyond the -1.9 percent abnormal return for union wins (the sum of the intercept and the union win coefficient) in firms organized by unions other than the five unions examined individually. While this incremental effect is large, the coefficient is not statistically different from zero at standard confidence levels (t-statistic $= -1.47$ percent).

Table 5 presents regression results for the two-digit SIC industries that most frequently appear in the sample: chemicals, machinery, electrical, and transport equipment. The results for industry regressions in table 5 are similar to the union regressions in table 4. In table 5, the intercept is significant for the petition month abnormal returns and insignificant for the combined petition and certification month abnormal returns. The union win coefficient is insignificant in both regressions. There is wide variation in the specific industry coefficients, and the only statistically significant coefficient is for a union win in the transportation industry. However, there are only two observations in this category, so the measured effect is suspect.

The results presented in tables 4 and 5 suggest that union and industry classifications are not sufficient to characterize the cross-sectional distribution of abnormal returns. While the coefficient estimates are suggestive, the corresponding standard errors are too large to make any inferences. Future research that focuses on the relation between abnormal returns and characteristics of unions and industries may provide an explanation for the differing effects of unionization.

TABLE 5

EFFECTS OF UNION WINS AND LOSSES FOR VARIOUS INDUSTRIES ON THE PETITION AND COMBINED PETITION AND CERTIFICATION DATE ABNORMAL RETURNS (*t*-Statistics in Parentheses)

$$Y = a_0 + \sum_{k=1}^{9} a_k D_{jk} + \epsilon_j$$

Independent Variables	Petition Date Abnormal Returns (%)	Combined Petition and Certification Date Abnormal Returns*	N Observations
Constant	-1.91 (-3.10)	-1.27 (-1.36)	253
Union win	-.54 (-.50)	1.12 (-.66)	54
Chemical loss	2.21 (1.87)	1.35 (.74)	21
Machinery except electrical loss	2.12 (1.35)	-.90 (-.37)	19
Machinery except electrical win	-3.50 (-1.23)	-6.29 (-1.43)	6
Electrical equipment loss	-.11 (-.09)	-.13 (-.08)	11
Electrical equipment win	.03 (.13)	-2.89 (-.76)	10
Transportation equipment loss	2.33 (1.51)	2.80 (1.19)	21
Transportation equipment win	6.65 (1.57)	14.01 (2.06)	2
R^2	.06	.04	

NOTE.—The four industries that appear most frequently in the sample are identified. The binary variables are defined as follows: $D_1 = 1$ for union wins and zero otherwise; D_2 [illegible] loss in the chemical industry (SIC code 28), $D_3 = 1$ for union losses in [illegible] otherwise, $D_4 = 1$ for union wins in the mach[illegible] losses in the electrical equipment (SIC code 36) [illegible] equipment industry and zero otherwise, D_7 [illegible] industry and zero otherwise, $D_8 = 1$ for un[illegible] regressions are estimated using weighted least s[illegible] the dependent variable.

* The abnormal return is measured as the su[illegible]

VI. Summary and Conclusions

This study examines the effect of unionization on profitability. Instead of measuring this effect through wage and production functions, the unanticipated changes in equity values are used as a measure of the effect of unionization on profitability. The rationale for using stock market data is threefold. First, since stock price changes measure the effect of new information on the present value of future cash flows, the impact of unionization can be measured without specifying production functions or the length of time required to adjust factor inputs. Second, the unanticipated change in equity values provides a complete measure of the effect of unionization including the impacts of changes in wages, productivity, and product prices. Third, the effects can be aggregated across firms and industries.

We examine the unanticipated changes in stock prices associated with 253 representation elections over the period 1961–80. Our empirical results show that stock prices fall when unions petition for representation elections. The average abnormal return is −1.38 percent, which is statistically significant with a *t*-statistic of −3.68. Since petitions raise the probability of eventual unionization, the significant negative abnormal returns at the time of petition indicate that unions are associated with reduced profitability. In the 54 instances in which the unions are successful, the combined abnormal return on the petition and certification months is −3.84 percent on average. In the 199 elections in which the unions are unsuccessful, stock prices decline by 1.32 percent. These value changes correspond to losses of $46,800 per worker for successful organization attempts and $7,400 per worker for unsuccessful organization attempts. While the difference between the abnormal returns for successful and unsuccessful unions is not statistically significant, the percentage losses for successful unions are three times larger than the losses for unsuccessful unions. This pattern is consistent with the hypothesis that unionization is associated with lower profits and reinforces the petition month results.

References

Allen, Steven G. "Unionization and Productivity in Office Building and School Construction." Mimeographed. Raleigh: North Carolina State Univ., Dept. Econ. and Bus., July 1982.

Brown, Charles, and Medoff, James. "Trade Unions in the Production Process." *J.P.E.* 86 (June 1978): 355–78.

Clark, Kim B. "The Impact of Unionization on Productivity: A Case Study." *Indus. and Labor Relations Rev.* 33 (July 1980): 451–69.

Fama, Eugene F. *Foundations of Finance.* New York: Basic, 1976.

Fama, Eugene F.; Fisher, Lawrence; Jensen, Michael C.; and Roll, Richard. "The Adjustment of Stock Prices to New Information." *Internat. Econ. Rev.* (February 1969): 1–21.

Freeman, Richard B. "The Exit-Voice Tradeoff in the Labor Market: Unionism, Job Tenure, Quits, and Separations." *Q.J.E.* 94 (June 1980): 643–73.

Freeman, Richard B., and Medoff, James. "The Two Faces of Unionism." *Public Interest*, no. 57 (Fall 1979), pp. 69–73.

Getman, Julius G.; Goldberg, Stephen B.; and Herman, Jeanne B. *Union Representation Elections: Law and Reality.* New York: Russell Sage Found., 1976.

Gifford, Courtney D., ed. *Directory of U.S. Labor Organizations 1982–1983.* Washington: Bur. Nat. Affairs, 1982.

Lewis, H. Gregg. *Union Relative Wage Effects. A Survey.* Draft manuscript. Durham, N.C.: Duke Univ., 1981.

National Labor Relations Board. *Annual Report* Washington: Government Printing Office, various years.

Neumann, George R. "The Predictability of Strikes. Evidence from the Stock Market." *Indus. and Labor Relations Rev* 33 (July 1980): 525–35.

Roomkin, Myron, and Block, Richard N. "Case Processing Time and the Outcome of Representation Elections: Some Empirical Evidence." *Univ Illinois Law Rev.* 1 (1981)· 1175–97.

Schwert, G. William. "Using Financial Data to Measure Effects of Regulation." *J Law & Econ* 24 (April 1981)· 121–58.

Thompson, Earl A. "On Labor's Right to Strike." *Econ Inquiry* 18 (October 1980): 640–53.

Comments

A Note on Extraction with Nonconvex Costs

Sheldon Kimmel

U.S. Department of Justice

I. Introduction

In a recent article, Eswaran, Lewis, and Heaps (1983) (ELH) model a market for an exhaustible resource in which each firm owns R units of the resource and can extract at a rate of q units for a cost of $C(q)$. $C(0) = 0$ and average extraction costs, $C(q)/q$, are assumed to exhibit economies of scale initially and reach a minimum at $q = \bar{q}$. Marginal cost is always positive, but it decreases until $q = \tilde{q}$ and then increases (i.e., total cost is not convex). The resource cannot be stored once it has been extracted. Demand does not change over time; if extraction proceeds at a rate Q at time t, then price is $P(Q)$ at time t. All firms are assumed to know the entire time path of prices at which the resource will be sold (time is taken to be continuous). Firms are assumed to be price takers in the sense that no firm can influence the time path of prices. ELH conclude that this model has no equilibrium. However, the model contains some assumptions that are not reasonable. This note relaxes those unreasonable assumptions, adds a simplifying assumption, and produces an equilibrium.

The conclusion that there is no equilibrium follows from two of ELH's most basic conclusions: no firm will ever extract at a rate be-

---

Lester Telser, Robert McGuckin, James Heckman, Tracy Lewis, Tim Brennan, and Gregory Werden made helpful comments on a previous draft. Any remaining errors are mine. The views expressed in this paper do not necessarily reflect the views of the Department of Justice

[*Journal of Political Economy*, 1984, vol. 92, no. 6]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9206-0005$01.50

tween zero and $\tilde{q}$,[1] and all firms act alike. Given these two basic conclusions, the lack of an equilibrium is easy to see. Since the extraction rate cannot be $\tilde{q}$ or more for longer than $R/\tilde{q}$ periods, there is a finite time t^* where no firms extract and each firm extracts at least $\tilde{q}$ just before t^*. This cannot be an equilibrium, since any firm would receive a higher value if it slightly delayed its last extraction. ELH claim that relaxing the price-taking assumption may not produce an equilibrium, but they do not support that claim.[2]

The problem with ELH's model is that it contains some standard assumptions that are reasonable when costs are convex but are not reasonable in a model with the cost function they assume. For example, ELH's firms are price takers. However, the number of firms is finite and the output of each firm is assumed to be continuous and nonnegligible while it is producing. Therefore, if even a single firm exits "industry output, and hence price, must jump discontinuously at the time firm j stops producing" (ELH, p. 166). Such a jump in output is quite reasonable given the initial scale economies in extraction. However, the ability of a single firm to cause price to jump in ELH's model is not consistent with the price-taking firms they assume. You cannot have a finite number of firms that extract continuously, enter and exit discontinuously, and are price takers.[3]

ELH's conclusion that all firms act identically is also a problem in their model.[4] To see the problems this causes, suppose that each firm's reserves were just sufficient to extract $\bar{q}$ for 1 period. Then whether we have a model with 10 firms and total reserves of $10\bar{q}$ or a model with 1,000 firms and total reserves of $1{,}000\bar{q}$, ELH's assumption forces the resource to be used up in 1 period. Surely if there are enough firms it becomes reasonable to have some extract initially and others wait and extract later. As with the price-taking assumption, this problem is relevant because of the initial economies of scale in extraction (i.e., the nonconvexity that gives us a minimum scale, $\bar{q}$). This problem is somewhat related to the previous problem, because the price-taking assumption is related to the identical behavior of the firms. In a model where firms do not behave identically, it is possible that many firms extract initially but only a few firms extract later on. The assumption of price-taking firms is not plausible in periods where only a few firms have any reserves left.

---

[1] Actually, price-taking firms would not choose to extract less than $\bar{q}$, which is a slightly stronger statement than ELH make since $\bar{q} \geq \tilde{q}$. To see this, let A be the set of times when extraction is positive but less than $\bar{q}$. ELH's adjustment-cost assumption rules out the possibility that A can be a set of isolated points. The measure of the set A is $\mu(A)$. Let total extraction on A be E. $E < \bar{q}\mu(A)$. Define a subset B of A of measure $E/\bar{q}$ such that $t \in B$ and $s \in A - B$ implies that $e^{-\delta t}[P_t\bar{q} - C(\bar{q})] \geq e^{-\delta s}[P_s\bar{q} - C(\bar{q})]$. Since $\bar{q}$ minimizes average cost, $P_t\bar{q}$ [illegible] $\in A - B$ imply that $e^{-\delta t}\{P_t - [C(q_t$ [illegible] extraction on A except on B where [illegible] extraction rates has no effect on [illegible] $- [C(q_s)/q_s]\}$ per unit. The gain [illegible] $e^{-\delta t}(\bar{q} - q_t)^{-1}\{[P_t\bar{q} - C(\bar{q})] - [P$ [illegible] $e^{-\delta t}\{P_t - [C(\bar{q})/\bar{q}]\}$ (the inequality [illegible] $t \in B$, that gain is at least $e^{-\delta}\{P$ [illegible] variation is profitable and extraction [illegible] that the distinction between $\tilde{q}$ and q is [illegible] cost becomes completely inelastic at q the [illegible]

[2] ELH also claim that "it seems intuitively clear that competitive [illegible] exist and that production will be inefficient if their cost structure was changed to one with fixed costs and constant marginal costs of extraction" (ELH, p. 155). Clearly with such a cost structure it would be inefficient to have more than one firm extracting at any point in time. If we demand efficiency we cannot have price taking firms. However, it seems likely that a Cournot-Nash equilibrium exists.

[3] This paper relaxes the assumption that firms are price takers. Mumy (1984) relaxes the assumption that there are only a finite number of firms (i.e., he takes the number of firms as continuous, as in a model with an uncountable set of firms of measure N). Mumy also shows by an example in discrete time that an equilibrium can result if there are a finite number of firms.

The interesting complication (nonconvexity) that ELH introduced into their analysis is the source of their difficulties. Assumptions that are reasonable in a world of convex costs simply cannot be transferred to a world where costs are not convex. When we introduce nonconvexities in a model with a finite number of firms we have to allow discontinuous price paths, firms that do not all act alike, and firms that realize that their decisions can affect prices. These complications make the control-theory approach that ELH took much more difficult. Unlike the simple control-theory model, the instantaneous profit function and the values chosen for the control variable are not continuous functions of time.

Since the goal of this paper is to show that there can be an equilibrium in ELH's world with more appropriate behavioral assumptions, it is more instructive to simplify the cost structure than it is to complicate the analysis. If we consider an extreme case of ELH's cost structure, then control theory is unnecessary and the essential features of the model become clear. In this note, as in ELH's paper, there is a positive rate of extraction, $\bar{q}$, which no extracting firm will extract less than (some support for this assumption appears as n. 1). The simplifying assumption is that marginal cost is completely inelastic for production of higher levels. Therefore a firm extracts at a rate $\bar{q}$ or not at all. In addition to introducing this simplifying assumption, this note relaxes ELH's questionable assumptions; price is not assumed to be a continuous function of time, firms are not assumed to act alike or to be price takers.[5] Instead, we assume that each firm takes the extraction of its competitors as given. Section II of this comment is a demonstration that, with these modifications, the model has a Cournot-Nash equilibrium. Conclusions appear in Section III. The Appendix presents the relatively simple two-firm case. In that case the social optimum is not a Cournot-Nash equilibrium.

---

[4] In their appendix, ELH argue that all firms act identically. That argument requires firms to be price takers but includes the sentence quoted in the previous paragraph. It has not yet been proven that all firms act identically in ELH's model.

[5] If we did not force q_t to be either zero or $\bar{q}$ we would also want to relax ELH's assumption that firms that are extracting act continuously. In that case, firms that

II. Equilibrium in the Modified Model

Each firm's problem is to choose q_t and T to

$$\max \int_0^T e^{-\delta t}[P_t q_t - C(q_t)]dt, \tag{1}$$

subject to $R(t) \geq 0$, $\dot{R}(t) = -q_t$, and $q_t \in \{0, \bar{q}\}$. This is identical to ELH's equation 1 except for the constraint on q_t added here to simplify the cost structure.[6] The symbol C will be used to stand for $C(\bar{q})$ to simplify the notation. The symbol P_t should not be thought of as price as a function of time, but rather as price as a function of the total level of extraction that will occur at time t if the given firm extracts q_t.

Only time paths of total extraction that are monotonic decreasing step functions with all steps of size $\bar{q}$ can possibly be equilibria in this model. All other time paths involve either increases in extraction or decreases of $2\bar{q}$ or more. An increase in extraction can happen only if at least one firm begins extracting at some time $t > 0$ even though it could have begun extracting a little earlier and faced the same number of competitors (or fewer competitors). This could not be an equilibrium given a positive interest rate and a non-upward-sloping demand curve. A decrease in extraction of $2\bar{q}$ or more could happen only if two or more firms simultaneously stop extracting. Such a firm would be more profitable if it p
just after its rival stopped ex
extracting at time t, any one
poned some of its extractio
period just after t).

Suppose the number of fir
tonic decreasing step functio
never skips a step (the number of m
much larger than N). Let t_1 be the first time wh

---

remain in the market would increase their extraction rate discontinuously in reaction to exit. A firm's exit causes output of the remaining incumbents and therefore marginal cost to jump up. However, increased extraction by incumbents does not make up for the decreased extraction by the exiting firm, so total extraction declines and price jumps up. The increased extraction by each incumbent and the price rise balance, so that marginal net revenue does not jump (i.e., P and C' both jump up so that marginal net revenue does not jump).

[6] The subscript ELH used to identify individual firms has been dropped to simplify the notation.

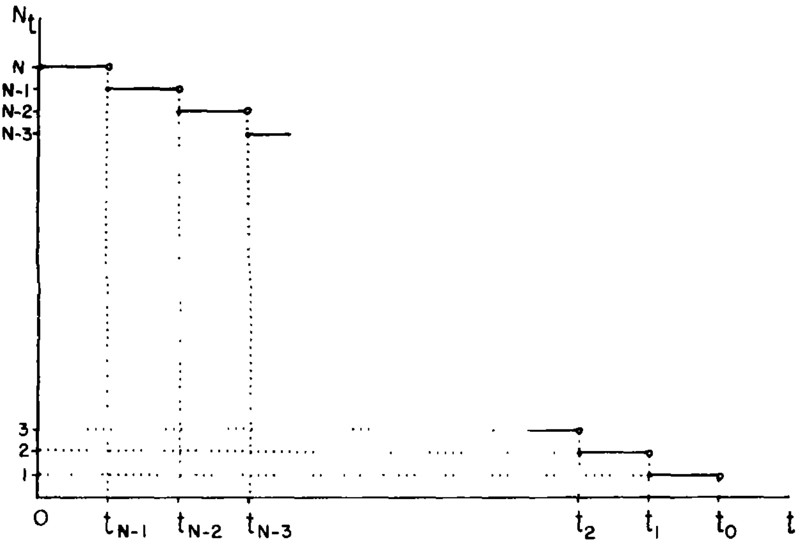


FIG. 1

($t_N = 0$, $t_0 = T$). If $t_i \leqslant t < t_{i-1}$, then the present value of the net revenue received by a firm extracting at time t is $I(t) = e^{-\delta t}[\tilde{q}P(i\tilde{q}) - C]$.

Define $E(t) = e^{-\delta t}\{\tilde{q}P[(i + 1)\tilde{q}] - C\}$, where $t_i \leqslant t < t_{i-1}$ (or $t_0 \leqslant t$ and $i = 0$). $E(t)$ is the present value of the net revenue a new entrant would get at time t. Suppose that the time path of extraction is such that, for any $i < N$, $E(t_i)$ is a constant independent of i and that this constant is between $E(0)$ and $I(0)$ (i.e., suppose the t_i's are chosen to arrange that). Call this constant W. It will be shown later (n. 7) that it is possible to choose t_i to arrange this.

Given any proposed time path of extraction for the firms, there are two types of variations that any particular firm might consider. First, it might consider extracting a total amount that is more or less than the total amount it would extract under the proposed time path. Second, it might consider switching the time at which it extracts but having a total extraction of the level set in the proposed time path (i.e., if the proposed time path has firm i extract zero at time t and $\tilde{q}$ at time s, the firm might consider extracting $\tilde{q}$ at time t and zero at time s). We deal with this second variation first.

PROPOSITION: If extraction proceeds as described above, then for each time t a firm extracting at time t has no incentive to switch its extraction to any other time.

PROOF: Both E and I are continuous and strictly decreasing functions of time except at times t_i. For $i < N$, $E(t_i) = W$ and for $i = N$, $E(t_N) \leqslant W$. $E(t)$ can attain a maximum only at some time t_i so W is the

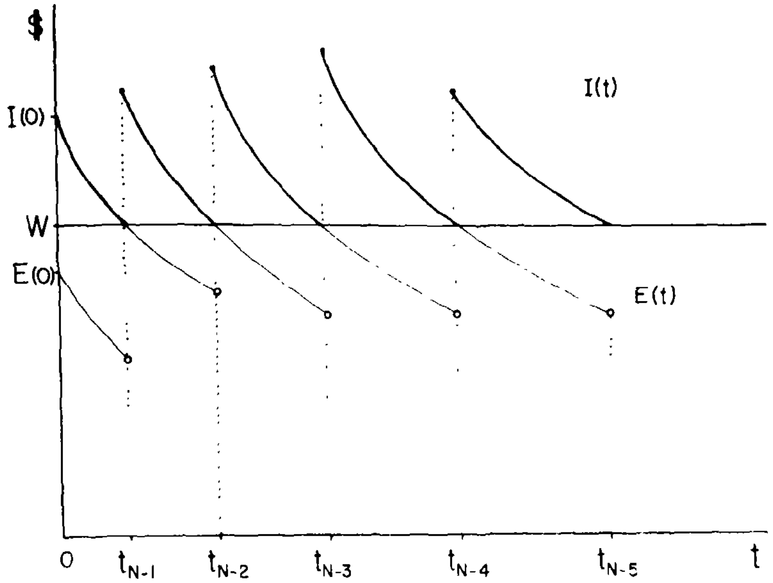


FIG. 2

maximum value of the function $E(t)$ (see fig. 2). For $i < N$, the limit of $I(t)$ as t increases to t_i is $E(t_i)$, since both functions describe situations with sales of $(i + 1)\bar{q}$ and, in the limit, the discount factors ($e^{-\delta t}$ and $e^{-\delta t_i}$) are the same. Since $E(t_i) = W$, $I(t)$ decreases toward W and then jumps up. Therefore, $I(t)$ is always greater than W. This means that a firm that was extracting at time t would [illegible] than W to get at
most W if it switched the ti
This switching would decre

COROLLARY: Given an inc
the number of firms extrac
step function that never sk
zero, and has discontinuiti

$$t_i = \frac{\ln\{\bar{q}P[(i + 1)\bar{q}]}{\delta}$$

then the firms would have no incentive to switch the times at which they extract.[7]

---

[7] We can now establish the earlier claim that t_i can be chosen in such a way that for $i < N$, $E(t_i)$ is a constant independent of i and is between $E(0)$ and $I(0)$. Given the demand curve, the cost function, and N, choose W to be any number between $\bar{q}P[(N + 1)\bar{q}] - C$ and $\bar{q}P(N\bar{q}) - C$. Now define t_i as in eq. (2) for $i < N$. Then $E(t_i)$ is W for $i < N$ and W is between $E(0)$ and $I(0)$. The equilibrium values of N and W are discussed below.

The next question to deal with is whether firms would want to extract the amount specified in the corollary. Presumably $\tilde{q}P(\tilde{q}) > C$, so firms would want to extract all their reserves. The total amount of extraction done if firms extract as in the corollary is

$$Q = \sum_{i=1}^{N} (t_{i-1} - t_i)i\tilde{q}, \tag{3}$$

$$t_{i-1} - t_i = \frac{\ln[\tilde{q}P(i\tilde{q}) - C] - \ln\{\tilde{q}P[(i+1)\tilde{q}] - C\}}{\delta} \quad \text{for } i < N, \tag{4}$$

$$t_{N-1} - t_N = \frac{\ln[\tilde{q}P(N\tilde{q}) - C] - \ln W}{\delta}. \tag{5}$$

For $i < N$, $t_{i-1} - t_i$ depends on the parameters of supply and demand, on i, and on δ, but not on N or W. Given N, W is defined in the corollary as being between $E(0)$ and $I(0)$. By varying N and W, any level of total extraction can be achieved with extraction proceeding as in the corollary.[8] To see this we must consider different values of N, so let superscripts on E or I denote the number of firms that initially extract. This allows us to rewrite equations (4) and (5) as follows:

$$t_{i-1} - t_i = \frac{\ln I^i(0) - \ln E^i(0)}{\delta} \quad \text{for } i < N \tag{4'}$$

$$t_{N-1} - t_N = \frac{\ln I^N(0) - \ln W}{\delta}. \tag{5'}$$

Let Δt_i be the value of $t_{i-1} - t_i$ from equation (4′). Let $Q(x, y)$ be the total extraction if $N = x$ and $W = y$. Then

$$Q[x + 1, I^{x+1}(0)] = \sum_{i=1}^{x} (\Delta t_i)i\tilde{q}, \tag{6}$$

$$Q[x, E^x(0)] = \left[\sum_{i=1}^{x-1} (\Delta t_i)i\tilde{q}\right] + \frac{x\tilde{q}\{\ln[I^x(0)] - \ln[E^x(0)]\}}{\delta}, \tag{7}$$

$$Q[x + 1, I^{x+1}(0)] - Q[x, E^x(0)] = \frac{x\tilde{q}\{\Delta t_x\delta - \ln[I^x(0)] + \ln[E^x(0)]\}}{\delta}. \tag{8}$$

---

[8] There is a minor complication. If the number of firms that were endowed with reserves would not be able to extract all of the reserves if they all began extracting as in the corollary, then the actual equilibrium is a corner solution: all firms extract until the level of reserves has fallen to the point where the number of firms can extract all of the reserves by the path described in the corollary, i.e., W will have to be small enough so that $t_{N-1} - t_N$ in eq. (5) is large enough—this will require $W < E(0)$. After that point, the number of firms extracting declines one at a time to zero. Equation (4) still holds for i smaller than the total number of firms.

Equations (4′) and (8) show that $Q[x+1, I^{x+1}(0)] = Q[x, E^{x}(0)]$. Given $N = x$, $W = I^{x}(0)$, we can continuously lower W until $W = E^{x}(0)$. This continuously raises Q to $Q[x+1, I^{x+1}(0)]$. At that point we can raise N to $x+1$ and continue lowering W. As we continuously lower W and raise N appropriately, Q rises continuously. Let N^* be the number of firms that would extract all of the reserves if they extracted initially and extraction proceeded as in the corollary.[9]

If N^* firms begin extracting and if the number of firms extracting decreases by one at each time t, (there may, at any time, be many firms stopping and starting extraction), then a price path will be generated such that no firm will have an incentive to switch extraction from one point in time to another. If W is chosen so that the total extracted equals the initial endowment of reserves, then no firm would have an incentive to vary its total extraction. This is an equilibrium.

III. Conclusion

Simplifying ELH's cost structure and relaxing the standard assumptions that are not reasonable in a model with nonconvex extraction cost (price-taking firms that all act alike and continuous prices) yields an equilibrium. Relaxing these assumptions without simplifying the cost structure gives us a model where the instantaneous profit function and extraction rates are not continuous. This makes the model more complicated than either the model in this note or the model in ELH's paper. However, it seems likely that the equilibrium when marginal cost becomes completely inelastic is similar to an equilibrium that exists in a general model with U-shaped costs. Although the time path of prices would still be discontinu[illegible] would be one difference betw[illegible] with U-shaped costs and th[illegible] model with U-shaped costs, p[illegible] so does extraction per firm b[illegible] net revenue does not jump.

This paper has simplified [illegible] their assumptions to get an equ[illegible] is expositionally helpful but does not [illegible] point of this paper is that relaxing their assumptions in a way that makes them more reasonable produces an equilibrium.

---

[9] It is not necessary to specify the extraction rate for each firm, but that might be fairly simple. For example, if no firms exhaust their reserves at t, then no firms enter and one incumbent firm exits (to reenter later when prices are higher and some firm has exhausted its reserves). If some firms exhaust their reserves at t, then they are the only firms that exit and all but one of them are replaced by new entrants. If t is not a t then the number of entrants at t equals the number of firms who exhaust their reserves at t.

Appendix

The Two-Firm Case

Equilibrium Extraction

It may be helpful to consider the simple case where there are only two firms. If $\bar{q}P(2\bar{q}) < C$, *then any feasible pattern of extraction where the two firms* never extract simultaneously, but total extraction proceeds at a constant rate until reserves are exhausted, is an equilibrium, and those are the only equilibria. If $\bar{q}P(2\bar{q}) > C$, then there are two cases to consider, depending on whether or not $R < (\bar{q}/2\delta)\{\ln[\bar{q}P(\bar{q}) - C] - \ln[\bar{q}P(2\bar{q}) - C]\}$.

In the case where reserves are small, all of the production patterns that are equilibria in the case where $\bar{q}P(2\bar{q}) < C$ are still equilibria. To see this, note that switching $\bar{q}$ units of reserves from a period t when you have been extracting to a period s when your competitor is extracting increases your profits by $\Delta\pi = e^{-\delta s}[\bar{q}P(2\bar{q}) - C] - e^{-\delta t}[\bar{q}P(\bar{q}) - C]$. Since $P(2\bar{q}) < P(\bar{q})$, $\Delta\pi < 0$ unless $s < t$. In fact the most profitable switch possible is the one with $s = 0$ and $t = t_0 = 2R/\bar{q}$. In that case $\Delta\pi$ is an increasing function of R. Therefore, by assumption $\Delta\pi$ is less than it would be in the case $R = (\bar{q}/2\delta)\{\ln[\bar{q}P(\bar{q}) - C] - \ln[\bar{q}P(2\bar{q}) - C]\}$. In that case $t_0 = \{\ln[\bar{q}P(\bar{q}) - C] - \ln[\bar{q}P(2\bar{q}) - C]\}/\delta$ so $\Delta\pi = 0$. Therefore, $\Delta\pi < 0$ in this case where R is small enough. Switching can only reduce profits, so the firms would not switch; they are in equilibrium.

In the case where reserves are large, suppose both firms extract until $t_1 = (R/\bar{q}) - \{\ln[\bar{q}P(\bar{q}) - C] - \ln[\bar{q}P(2\bar{q}) - C]\}/2\delta$. The reader should note that $t_1 > 0$ by the assumption that R is large. After t_1 one firm or the other (but not both) extracts until $t_0 = (2R/\bar{q}) - t_1$.[10] In that case total extraction is $2\bar{q}t_1 + \bar{q}(t_0 - t_1) = 2\bar{q}t_1 + \bar{q}[(2R/\bar{q}) - 2t_1] = 2R$. This pattern of extraction is an equilibrium if no firm would switch extraction from a period t when it was extracting to a period s when it was not extracting. There are two cases to consider: $t < t_1$ or $t \geq t_1$. If $t < t_1$, then the firm receives $P(2\bar{q})$ at time t. The present value of what it loses by switching is smallest for $t = t_1 - \epsilon$. The switch is unprofitable for $s < t_0$ since the price that an entrant gets then is $P(2\bar{q})$, so switching just delays the receipt without affecting its size. The most profitable choice for s is therefore $s = t_0$. The other case to consider is if $t \geq t_1$. In that case the firm receives $P(\bar{q})$ at time t. The firm cannot switch to an s that gives it as good a price at an earlier date, and it can never get a better price. Therefore, the most profitable s it might be able to switch to is $s = t_1$. This involves the greatest increase in profit if $t = t_0 - \epsilon$. Therefore, if it can be shown that neither switching from $t_1 - \epsilon$ to t_0 nor switching from $t_0 - \epsilon$ to t_1 is profitable, then we can conclude that no switching is profitable, and we have an equilibrium. In both cases switching increases profits by $\Delta\pi = e^{-\delta t_1}[\bar{q}P(2\bar{q}) - C](1 - e^{\delta\epsilon}) < 0$. This can be easily shown using the fact that $e^{-\delta(t_0 - t_1)} = [\bar{q}P(2\bar{q}) - C]/[\bar{q}P(\bar{q}) - C]$.

If there are only two firms, then there are essentially only two cases. If $\bar{q}P(2\bar{q}) < C$ or R is small enough, then any feasible pattern of extraction where *the two firms never extract simultaneously but one or the other firm is always* extracting until the reserves are exhausted is an equilibrium. If $\bar{q}P(2\bar{q}) > C$ and R is large enough, then one equilibrium involves both firms' extracting until t_1 (as defined above) and then any feasible pattern of extraction where the two firms never extract simultaneously but one or the other firm is always extracting until the reserves are exhausted.

---

[10] The values for t_0 and t_1 were constructed using eq. (4) and the assumption that extraction exactly exhausts reserves.

Optimal Extraction

Society gets a total surplus of $V_1 - C$ from extraction of $\bar{q}$ units and $V_2 - 2C$ from extraction of $2\bar{q}$ units. Since the discount rate is positive, the optimal extraction path consists of extraction at a rate of $\bar{q}$ or $2\bar{q}$ until reserves are exhausted (i.e., there is no point in having extraction of zero followed by positive extraction). If it is optimal to set $q_t = 2\bar{q}$, then it is optimal to extract at a rate of $2\bar{q}$ for all times before t.[11] Therefore, the only possible optimum extraction paths involve extraction at a rate of $2\bar{q}$ until t_1 and then extraction at a rate of $\bar{q}$ until reserves are exhausted at t_0.

The reserves can be extracted in two ways. Either one firm extracts until $R/\bar{q}$ and then the other firm extracts until $2R/\bar{q}$ (in which case $t_1 = 0$, $t_0 = 2R/\bar{q}$), or both firms extract until t_1, then one firm temporarily stops extracting until the other firm's reserves are exhausted (i.e., until $R/\bar{q}$), at which point the firm reenters extraction until $(2R/\bar{q}) - t_1$. In this section we assume that R is so large that $V_2 - V_1 - C > (V_1 - C)e^{-2\delta R/\bar{q}}$. In this case it will be optimal for both firms to extract initially. The total surplus V is given in equation (A1):

$$V = \int_0^{t_1} (V_2 - 2C)e^{-\delta t}dt + \int_{t_1}^{(2R/\bar{q})-t_1} (V_1 - C)e^{-\delta t}dt. \tag{A1}$$

This is maximized by choosing t_1 to be $t^* = (R/\bar{q}) - [\ln(V_1 - C) - \ln(V_2 - V_1 - C)]/2\delta$. This completely describes the optimal rate of extraction. This optimal value of t_1, t^*, is an equilibrium if and only if $[\bar{q}P(\bar{q}) - C]/(V_1 - C) = [\bar{q}P(2\bar{q}) - C]/(V_2 - V_1 - C)$. In words, the optimum is an equilibrium if and only if the ratio of a firm's net revenue to its contribution to total surplus is invariant with respect to total production. This may be true for some demand curves, but it is not true in general (although it would be true if the firms could price discriminate perfectly).

References

Eswaran, Mukesh; Lewis, Tracy
of Market Equilibria in Exh
Costs." *J P E* 91 (February 1
Mumy, Gene E. "Competitive
with Decreasing Costs. A Con
onstration of Nonexistence

---

[11] PROOF Consider an optimal extraction path with $q_t = 2\bar{q}$, $q_s = \bar{q}$. We could vary this path by changing q_t to $\bar{q}$ and q_s to $2\bar{q}$. That increases the present value of the surplus by $e^{-\delta s}(V_2 - 2C - V_1 + C) + e^{-\delta t}(V_1 - C - V_2 + 2C) = (e^{-\delta s} - e^{-\delta t})(V_2 - V_1 - C)$. This is negative (i.e., the path is optimal) if and only if $V_2 - V_1 - C$ and $t - s$ have different signs. Clearly $V_2 - V_1 - C$ is not negative, since it would not be optimal to extract $2\bar{q}$ at any time in that case (i.e., it would be preferable to throw the resource away). Since $V_2 - V_1 - C > 0$, $s \geq t$ on the optimal path we considered. Since $q_s = \bar{q}$ implies that $s \geq t$, we conclude that $s < t$ implies $q_s \neq \bar{q}$ (i.e., $q_s = 2\bar{q}$).

Competitive Equilibria in Exhaustible Resource Markets with Decreasing Costs: A Comment on Eswaran, Lewis, and Heaps's Demonstration of Nonexistence

Gene E. Mumy

Ohio State University

In a recent article, Eswaran, Lewis, and Heaps (1983) (ELH) correctly pointed out that many analyses of exhaustible resource markets assume the existence of a competitive equilibrium for the industry without checking to see if such an equilibrium is consistent with the behavior of the industry's constituent mining firms. ELH have attempted to rectify this situation by analyzing the behavior of an industry's firms and seeing if that behavior is consistent with the existence of a competitive equilibrium. ELH reach the startling conclusion that it takes nothing more than firms with U-shaped average cost curves, implying an initial range of increasing returns to scale, to eliminate the existence of a competitive equilibrium.

On its own terms, ELH's analysis is sound and provides a needed corrective to a major oversight in the literature. However, given the dramatic nature of their conclusion, the assumptions on which ELH's analysis is based require careful scrutiny. On this score, it seems to me that ELH's assumptions are implicitly at odds with the conditions that economists usually think of as characterizing competitive price-taking behavior.

To show this, ELH's analysis is summarized, in Section I, so as to emphasize that their results depend crucially on the assumption that the number of producing firms cannot be treated as a continuously changing variable over time. In Section II, intertemporal equilibrium is characterized for the case where the number of producing firms is treated as a continuous variable, and it is argued that this assumed continuity is not only convenient but actually required for consistency with the assumption of competitive price-taking behavior. Then, to further demonstrate the dependence of the nonexistence result on continuity assumptions, an equilibrium solution is computed for a specific case set in discrete time. Section III contains conclusions.

[*Journal of Political Economy*, 1984, vol. 92, no. 6]
© 1984 by The University of Chicago. All rights reserved. 0022-3808/84/9206-0006$01.50

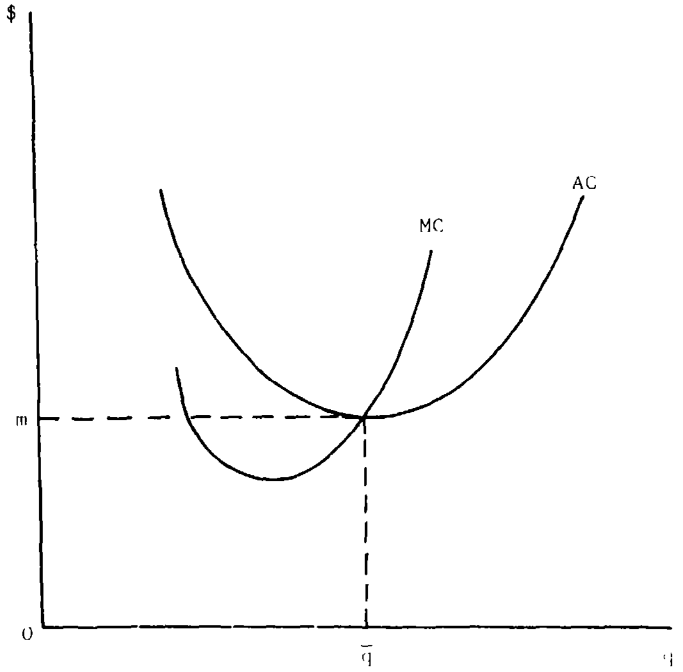


FIG. 1

I. ELH's Demonstration of Nonexistence

The basic case dealt with by ELH
number, N, of identical firms
of some resource. The ith fir
with the cost function bein
shaped average cost curve, sl
cost-minimizing rate of outp
which also equals marginal co
price function, $P = P(Q)$, which de
is characterized by $p' < 0$ and finite $P(0)$. The
with $Q(t) = \Sigma_{i=1}^{N} q_i(t)$. Last, ELH assume that all prices and quantities must be determined at the initial point in time by a market process with complete forward contracting, as this avoids possible problems of market power during the latter stages of industry production. Although ELH are not specific about this, it is useful to think of the market process as being a Walrasian *tâtonnement*. With these assumptions in place, ELH move to their demonstration that a competitive intertemporal equilibrium does not exist.

ELH state the crucial proposition that $q_i(t) = q(t) > 0$ for all i, over all t in the time interval $(0, T)$, which is the duration of industry production. This also means that $n(t) = N$ for all t in the interval $(0, T)$, where $n(t)$ is the number of firms producing at time t. ELH relegate the proof of this proposition to the appendix and proceed with their demonstration, which is quite simple. They first derive the standard result that each firm will choose an output trajectory, $q(t)$, such that the price net of marginal cost increases at the rate of discount; that is, $(d/dt)[P - C'(q)] = r[P - C'(q)]$, where r is the discount rate. Now if all N firms produce the same $q(t)$, given the static demand function, the only way net price can increase is for each of the firms to move back down the upward-sloping part of its marginal cost curve, arriving at $q(T) = \bar{q}$ in the terminal period, as is correctly demonstrated by ELH. But the problem for ELH is that $P(T) = P(N\bar{q})$, which is less than $P(0)$, the latter requiring that $Q(T)$ be zero. Therefore, at $t = T$ price jumps discontinuously, meaning all firms have an incentive to defer some production to take advantage of the higher price. But this contradicts the behavior that causes all firms to locate $\bar{q}$ units of output in the same terminal period T. Ergo, equilibrium does not exist. This case is illustrated by the curve NN in figure 2. A

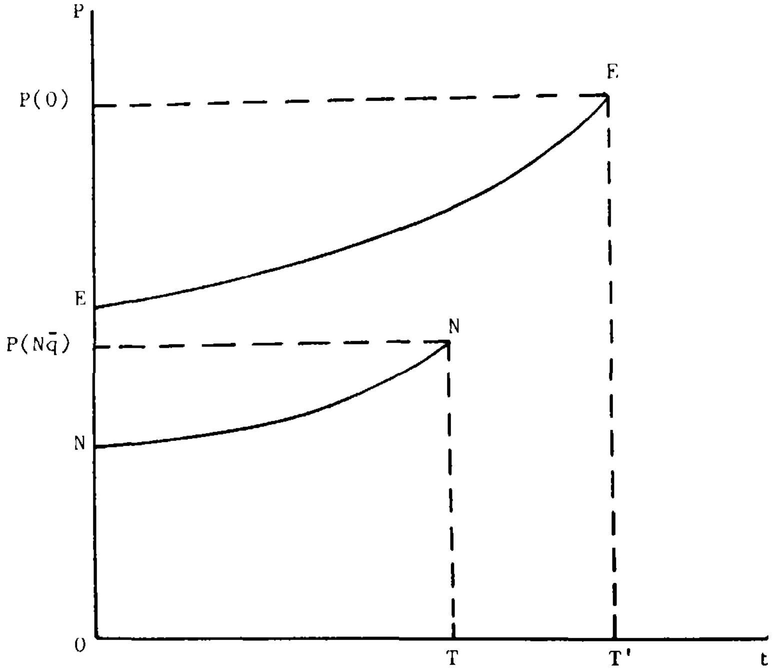


FIG. 2

moment's reflection, however, suggests that this cause of nonexistence would not be a problem if firms would choose to produce during different subintervals of the industry's life and did not all choose the same terminal period. The nonexistence result depends crucially on ELH's proof that all firms have identical production profiles, and this, in turn, requires that the number of producing firms not vary over time.

The essence of the ELH proof is that no firm can find it profitable to defer production from an earlier point in time to a later point in time if it is profitable for any firm to be operating at the earlier point. Their reasoning is that postponing production can be profitable only if the market price is increasing. But, given the static demand function, we can have only $\dot{p}(t) = (dp/dQ)\dot{Q}(t) > 0$, if $\dot{Q}(t) < 0$. This, they argue, must imply that there are some firms at the earlier date for which $\dot{q}(t) < 0$, meaning, since $q(t)$ is continuous, that these firms are moving down their marginal cost curves. Whenever this is occurring, however, it is profitable for all firms to be producing in that earlier period, so firms will never postpone production during periods when other firms are producing. The question, though, is, Does a continuous increase in $P(t)$ require a continuous decrease in $q(t)$ for some firms?

If we agree with ELH that all firms producing at a given t will be producing at identical output rates $q(t)$, then $Q(t) = n(t)q(t)$. In this case, $\dot{Q}(t) < 0$ does not require firms to be moving along continuously tilted production profiles. Rather, if firms have flat production profiles, when they are producing, $\dot{Q}(t) < 0$ can be generated with $\dot{n}(t) < 0$ and $\dot{q}(t) = 0$. It is precisely this variation in the number of producing firms that allows a compet[illegible] the market process.

II. Competitive Equilibrium

Imagine a situation where fir[illegible] the average cost-minimizing r[illegible] $= c'(\bar{q}) = m$. Also suppose that va[illegible] production is achieved by varying the number [illegible] The question is, Does a price trajectory exist such that this situation can be sustained as a market equilibrium? The answer is yes.

The general form of an equilibrium price trajectory is $p(t) = (p_0 - m)e^{rt} + m$, with $p(T) = P(Q)|Q = 0$ and initial price, p_0, determined by total reserves. With $\bar{q}$ as the production rate by producing firms, this trajectory is generated by a decrease in the number of producing firms over time, going to zero at $t = T$. This is an equilibrium because price net of marginal cost (equals m) rises at the rate of discount. A firm is always indifferent between producing $q(t) = 0$ or $q(t) = \bar{q}$ as long as production can be transferred between periods along this price trajectory, so firms have no incentive to alter their production profiles. There is also no incentive to extend industry life because price rises to choke off demand just as the industry runs out of reserves. This equilibrium solution is shown by curve *EE* in figure 2.

ELH's implicit objection to the equilibrium solution presented above is that the number of firms is obviously an integer function, and therefore it is inappropriate to treat it as a continuous variable. But if this is so, then a discrete change in the number of firms producing $\bar{q}$ must result in a discrete change in Q, and this must imply a jump discontinuity in the price trajectory. This means that no firm would have an incentive to stop producing, thus contradicting the possibility that the number of firms decreases, in which case we are back to ELH's nonexistence result.

Assuming that $n(t)$ is a continuous variable is, of course, an extraordinarily convenient analytical simplification. But more than that, it is appropriate given an analysis based on continuity in all other respects. Refusing to grant the assumption that n is continuous either calls into question the assumed competitive nature of the market or leads to the conclusion that the assumed continuity of some of the other functions must be abandoned. As for the first possibility: if n's being discrete means that one firm's stopping production causes a discontinuous jump in price, which influences that firm's behavior, then denying that n can be treated as a continuous variable is tantamount to denying a competitive market. Assuming n to be continuous is part of assuming a competitive market. As for the second possibility: if we want to maintain that n is discrete and that the market is competitive, then we need an analytical structure that does not require continuity for the price and quantity trajectories. Most obviously, perhaps, a discrete-time and discounting model suggests itself.

The issues at hand are nicely illustrated by a market equilibrium solution to the following example set in discrete time. For a succession of market dates, with intervals of 1 period, the static demand-price function for a resource is given by $P = 26 - (0.1)Q$. There are 74 identical mining firms ($N = 74$) with the same U-shaped average cost curve for producing and trading output on successive market dates. The minimum average cost is \$5 ($m = 5$) obtained by producing 10 units per market date ($\bar{q} = 10$), which also happens to be the total reserves of each firm. The constant 1-period discount rate between market dates is $r = 1$. All prices and quantities are determined on market date $t = 0$ by a Walrasian auctioneer who attempts to find the price series that clears the market at $t = 0$ and all successive market dates ($t = 1, 2, 3, 4, \ldots$).

TABLE 1

INTERTEMPORAL EQUILIBRIUM

	MARKET DATE					
	$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$ and Greater
P	6	7	9	13	21	26
n	20	19	17	13	5	0
$\bar{q}$	10	10	10	10	10	.
Q^S	200	190	170	130	50	0
Q^D	200	190	170	130	50	0

Suppose the auctioneer presents the market with the price series shown in the first row of table 1. In response to this price series each firm chooses to produce its entire reserves, $\bar{q}$, for a single market date Suppose that the number of firms, n, that choose to produce for each market date is as shown in the second row of table 1.[1] Then the total quantity supplied on each date is as shown in row 4 and quantity demanded, obtained by consulting the demand-price function, is as shown in row 5. What we have here, of course, is a competitive equilibrium.

In this case, net price increases at the rate of discount until the industry runs out of reserves, that is, $P_t - m = (p_0 - m)(1 + r)^t = 2^t$, so no firm has an incentive to change the date on which it sells its output. There is also no incentive to extend the industry's life because price chokes off demand before it can rise enough to cause any firm to defer production further. This last point also illustrates how the discrete case captures the essence [illegible] competitive market. The elim [illegible] $t = 4$ is not sufficient to ca [illegible] alters any firm's behavior. [illegible] and the auctioneer declares [illegible]

Quite obviously, the examp [illegible] and the equilibrium solution [illegible] ter values. Indeed, it is fairly [illegible] parameter values lead to cases where [illegible] Nevertheless, this is a clear counterexample to the claim that U-shaped average cost curves alone are sufficient to eliminate the existence of competitive intertemporal equilibria. When equilibrium does not exist, my suspicion is that U-shaped average cost curves, in conjunction with other parameters, make it inappropriate to describe the intertemporal resource market as characterized by competitive, price-taking behavior.

---

[1] Given identical firms there is no mechanism to determine which firms produce on which dates. The only contention, then, is that if the alignment shown in table 1 is somehow arrived at it will turn out to be an equilibrium solution

III. Conclusions

A competitive equilibrium, in an exhaustible resource market, may not exist for several reasons. It seems unlikely, however, that U-shaped average cost curves, by themselves, generate nonexistence. When firms do have U-shaped average cost curves and a competitive equilibrium does exist, it is characterized by firms with flat production profiles at the level of the average cost minimizing rate of output. On the other hand, total industry production varies over time because the number of producing firms also varies over time. In addition to being an equilibrium, this market solution also has the attractive characteristic that any rate of industry output is produced at minimum average cost.

Reference

Eswaran, Mukesh; Lewis, Tracy R.; and Heaps, Terry. "On the Nonexistence of Market Equilibria in Exhaustible Resource Markets with Decreasing Costs." *J.P.E.* 91 (February 1983): 154–67.

Book Reviews

The Laws of Human Relations. By HERMAN H. GOSSEN. Translated by RUDOLPH C. BLITZ.
Cambridge, Mass.: MIT Press, 1983. Pp. cxlix + 315. $47.50

Herman Heinrich Gossen and Richard Cantillon are surely leading contenders for Most Pathetic Tragedy Award in the history of economics. Gossen (1810–58) wrote a treatise presenting the fundamental theory of utility, *The Laws of Human Relations* (1854), which received absolutely no attention in the 4 years between its appearance and his early death. Cantillon (1685?–1734), a successful merchant for whom recognition of his *Essay on Commerce* was a less important matter—or he would have published it himself during his lifetime—was killed by a discharged French cook.

Gossen is difficult even for Germans to read, and the services of Professor Rudolph C. Blitz in translating it are above and beyond any call of duty. In addition to presenting a clear, accurate translation, he has lightened the task of the reader by dividing its unbroken text into chapters with headings. Readers need no longer feel that they are in a heavy forest, where they may encounter a wholly unexpected animal at any moment.

Professor Georgescu-Roegen has made the valuable contribution of a preface of monograph length about Gossen and the book. All the meager details of Gossen's life are assembled (a large fraction consisting of answers to civil service examinations!), the theory is reformulated and developed, and the
many negligent treatments or oversights of
sophistication often goes well bey
translation should be read befor

In addition to a prescient anal
is a protracted, often ingenious
economics. There is an extensive
treating present owners equitabl
a case for identical education of
other causes. A few fertile hints
modern distinction between luxuries
and the marginal productivity theory is touched upon
be recommended not only to those interested in the history of economic thought but also to economists.

GEORGE J. STIGLER

University of Chicago

[*Journal of Political Economy*, 1984, vol. 92, no. 6]
© 1984 by The University of Chicago. All rights reserved.

The Economics and Politics of Oil Price Regulation: Federal Policy in the Post Embargo Era. By JOSEPH P. KALT.
Cambridge, Mass.: MIT Press, 1981. Pp. 327. $45.00 (cloth); $12.50 (paper [1983]).

Conditions have changed rapidly in the energy sector. At the time this book was written, world oil prices were at their peak, having risen sevenfold in real terms over the previous 8 years. In this context, Kalt analyzed the struggles within the United States to moderate and redirect the associated rent transfers. As of mid-1984 the picture has changed substantially. The dollar price of world oil has fallen 35 percent below its peak, and the prevailing expectation is of stagnant or falling prices for the next decade. Gone are the oil price controls that are the focus of this book, with the federal Windfall Profits Tax and various state taxes left behind to collect some of the remaining rents. The policy conflict has shifted to natural gas, where there is another "The Economics and Politics of . . ." waiting to be written.

In such a circumstance, the Kalt study is like a contemporary analysis of the Civil War. He explains how the campaigns evolved over the years of conflict and attempts to assess who ultimately gained and lost. He also is concerned about how we got into such a costly affair in the first place. Two questions ought to be addressed to such a work: What does it contribute to our understanding of the oil policy wars of the 1970s? And what does it teach about analytic methods that might be applied to similar situations in the future?

On the first score, the Kalt book is the best single work available on the experience of oil price controls in the 1974–81 period. And a tangled story it is. When the first oil shock occurred there was a system of oil price controls already in place—a hangover from the Nixon anti-inflation scheme of 1971. They were modified and extended and used to hold down the price of domestic crude oil so that people downstream (oil refiners, distributors, and the ultimate consumers) got a lower average price of domestic-plus-imported supplies. To these crude oil price controls was added a system of "entitlements" to ensure that the cheap domestic crude oil (roughly 60 percent of total consumption at the time) was spread evenly over the refinery sector.

Given a political mandate to limit income transfers to producers it was an elegant solution. A shadow system of public finance, unique to the oil sector, was created—complete with taxes, transfers, and (no surprise) deadweight loss. In practice the system grew to mind-bending complexity as the various players (regions, consumers, refiners, and producers holding various classes of oil reserves) fought over the goodies. The book does a good job of sorting out the combatants, the evolving policies, and their effects, although it is not always easy going because there are so many details that matter.

Kalt uses an array of techniques—graphical, analytical, and empirical—to explore the economics of these schemes. Indeed, the use of a combination of methods and forms of presentation makes the book a valuable source of readings—say, for a course on the economics of energy and natural resources. Several issues are addressed, including the effect of the price controls on the average acquisition cost of crude oil to U.S. refiners, the associated changes in domestic supply and demand for products, and the ultimate distributional impact. A very contentious issue at the time (and the one to which Kalt devotes the most effort) was the question who actually benefited from the $15–$45 billion (depending on the year) producers were denied. In the mid-1970s there was a group of analysts who held that the oil price controls were a fraud on the consumer: U.S. product prices were set in world product markets (so the argument went), and there was no way for controls on crude oil to affect prices at the pump. The rents were being transferred to refiners in the form of increased margins.

It is a complex issue, requiring a sorting out of the influence of various control schemes (which themselves affected the degree of U.S. exposure to international markets) and different oil products. (The United States imported around half its residual fuel oil, but only small fractions of gasoline and middle distillates.) Kalt analyzes the question in three separate ways. He first goes through an exercise with rough estimates of supply and demand elasticities, to assess the likely interaction of U.S. and international markets. This is supplemented by the study of realized sales prices in the United States and overseas, as reflected in oil company accounts. Finally, he carries out an econometric analysis of *Platt's* data on refined product prices to explain why U.S. product prices should have fallen below European (Rotterdam) levels during the years when the domestic crude controls were binding. The results (e.g., that roughly 40 percent of the crude oil subsidy was ultimately realized in product prices) are convincing, as are the direction and general magnitude of the distributive consequences, compared with a hypothetical decade without controls.

There is uneasiness about the economic analysis at several points, and it arises where assumptions about market structure and behavior are made with insufficient attention to institutional detail. The book shows an impatience with possible impediments to a clean line of microeconomic argument. For example, early in the analysis it is assumed that the United States is a price taker in world crude markets, with citations to optimal-path cartel models as justification. At their best, these models offer insight into how a one- or two-person OPEC might operate in a world where only crude oil mattered and where there were no spot markets, no inventories, and no swings in expectations about future volume and price. In fact, U.S. market developments have a significant influence on OPEC's ability to carry out a price strategy, and it has been so throughout the period under study. Very likely Kalt's key results would be little changed by a more complex [illegible] issue is not discussed, so it is diffic[illegible]

Similarly, refiners are treated a[illegible] basis. There is a suggestion th[illegible] markup pricing reveal a "failur[illegible] input, output, and price determin[illegible]tional, but possibly wrong in spe[illegible] to try to figure out how the result[illegible] from this neat neoclassical world[illegible]

There is an extra ingredient in the K[illegible]nomic effects of federal policies to study the interest-group [illegible] pressures that led to them. The analysis focuses on 36 roll call votes on oil price-control issues in the Senate. Each vote is characterized as pro- or antideregulation, and econometric methods are used to try to sort out the influences. Senators are credited with two motives: "investment," which leads to support for measures that increase home-state wealth (thereby gaining votes), and "consumption," which encompasses everything else—most particularly expressions of political ideology or personal conceptions of the public good. Home-state economic stakes include the weight of the oil industry (production and refining) in the economy and interests of the state's consumers (represented by state expenditure on all energy). The proxy for ideology is the vote rating of the Americans for Democratic Action.

Taken as insight into the Senate votes per se, the analysis is well worth reading and the results are plausible. Both distribution and ideology are shown to matter. On the other hand, even on these narrow grounds there are problems. A state-by-state price index was available for all energy but not for the topic at hand, which is oil. As a result, the important political coalition of the Northeast and northern tier states (which are heavy residential and industrial users of oil products) is missing from the analysis. Also, in an era when national political action committees can raise (and target) large campaign chests, the very definition of a senator's "constituency" is more complex than this study can consider, and the two-way breakdown of motives is called into question.

Taken as a general conception of the politics of oil (as Kalt apparently intends), it is a breathtaking simplification of the legislative and regulatory process. One need only ask, What is the significance of a Senate floor vote on the overall path to new legislation—including executive agency politics, administration initiatives, House-Senate disputes, and the internal machinations of the many relevant Congressional committees? One applauds Kalt's effort for a willingness to probe this aspect: too many regulatory studies suffer from an insensitivity to the political side of the picture. But a main achievement of this attempt is to define clearly the *limits* to the amount of the action that can be captured with an economist's concepts and econometric tools.

That it failed to do more is but a small criticism of a highly competent piece of policy analysis, with important implications for future attempts to regulate the energy sector.

HENRY D. JACOBY

Massachusetts Institute of Technology

The Political Economy of U.S. Tariffs. An Empirical Analysis. By RÉAL P. LAVERGNE.
Toronto: Academic Press, 1983. Pp. 210. $24.50.

Because tariff legislation in the United States once provided, as the famous phrase puts it, a "paradise for pressure groups," and because the data and theory were available for the ready estimation of the incidence of the tariff, regression analysis of the cross-industry variation in tariff rates was one of the first empirical applications of the economic theory of economic regulation or rent seeking (Lavergne cites 24 such studies, 10 for the United States alone). The goal of this book is to improve our understanding of why some American industries, in selected years since 1930, received higher rates of import protection or had their protection cut by less than did others. The approach taken by Lavergne is to build some of the relevant institutional detail of the American tariff-making processes into his cross-sectional model, a commendable methodology in a piece of applied economic analysis. Lavergne presents the results of 11 cross-sectional regressions on 11 different dependent variables—nominal and effective tariff levels in 1964, 1972, and 1979; changes before these dates; measures of nontariff barriers—using a common set of 32 regressors; maximum sample size is 300. He displays a strong tendency to interpret the results of the 11 regressions collectively, so that a regressor and, more important, the theory behind its use are approved if the regressor frequently turns up with a large and significant coefficient. Thus no special attention is given to distinguishing explanations for the fall in tariff rates from those for the growth of nontariff measures, or to exploring connections between these phenomena. As much as to establish securely some detailed empirical results, the author wishes to advance certain kinds of explanations and to cast doubt on others. The lines are drawn early: "In contrast to the economistic model, the model developed in this book assumes that the behavior of politicians and bureaucrats is qualitatively very distinct from that of the consumer in the marketplace. The difference is that moral responsibility, as *reflected in the application of various norms and principles of decision-making*, comes to have a much more predominant role" (p. 12). Connected with this contrast is another, between an "equilibrium" or static approach and one that "does take into account the historical determinants of the tariff structure and the process of change whereby that structure changes over time" (p. 46). There is much to commend in the general aim of basing empirical models securely within their institutional and historical setting. What is, however, disputed below is Lavergne's acceptance of the "principled" model and decisive rejection of the "economistic" model

History is not bunk. The 1929/30 tariff was the last one passed by Congress. Since 1930, Congress has refrained from specifying through legislation the tariff rates to be applied to the various imports and has delegated to the executive the power to negotiate tariff reductions with other countries Until the formation of the GATT, negotiations were bilateral (although the most-favored-nation rule still applied). In the 1950s, the method was multilateral negotiations using the principal supplier approach, subject to something like bilateral balance of concessions The 1960s and 1970s saw flexible (or discretionary) application of tariff-cutting formulae. The whole process was operated within a broad set of "rules," most explicitly laid out in the escape clause first incorporated in the bilateral agreement with Mexico in 1942 and subsequently included in GATT, which allowed for the growth of imports into American markets but reserved to America the right to slow down that growth if an import or class of imports caused too severe a domestic political difficulty. (As strange as they may appear to many economists these general
rules have made possible the progress
extent not seen since the middle
extremely interesting that Laver
most important right-hand-side
American tariff rates for 1964, 1
clearly visible in contemporary ta
age tariff has fallen to about one

This, then, leads to one of
apologies for doing so, previous au
assumed that the tariff at time *t* was a fu
dent variables at or close to time *t*. In contrast, Lavergne posited an imp
degree of historical continuity, captured by the inclusion of the lagged endogenous and other variables

Connected but more important is Lavergne's criticism of previous work as being founded on the "artificial and basically incorrect assumption that people behave in fundamentally the same way in politics as they do in the market" (p. 31) The economistic model, central to the method used by adherents to the Tullockian approach to public choice or the Stiglerian approach to the economics of politics, asserts that, whether in public or private life, individuals attempt to maximize utility subject to constraints. It assumes the pursuit of self-interest by private individuals (through the ballot box, as well as campaign contributions, bribes, and other forms of pressure) and by politicians and bureaucrats; where self-interest is often but not accurately interpreted in strict terms of wealth. Explicitly rejecting a model in which demand for tariffs depends solely on pressure group activity and supply depends solely on maximization of personal profit by privileged politicians, Lavergne wishes to incorporate in his model the widespread perception or social expectation, which acts as a moral or legal constraint, that it is "the government's duty to serve the general public interest" (pp. 12–13).

Accepting, as I do, that such a perception is widespread, can we discover its systematic effects on public decisions such as those about tariff rates? There is a danger, to which Lavergne is not blind, that on discovering a significant regression coefficient or pattern of coefficients the researcher may rationalize the results as reflecting a moral or social consensus. For example, from 1934 on the primary announced goal of the American reciprocal trade agreements program has been to expand American exports, with increases in American imports seen as a cost. Because of the largely bilateral nature of the negotiations, it is not unexpected to find that tariffs in the 1960s and 1970s were lower on products of especial interest to exporters in countries where the United States had large potential markets. Lavergne's regressions show that where an import was heavily sourced from Canada it has a lower tariff, and where heavily from Japan and (other) developing countries, a higher tariff. His preferred interpretation is that these results give relatively unambiguous support to a model based on principled behavior, that is, on efforts by politicians and bureaucrats to maximize social welfare (pp. 132–35) However, in his conclusions (p. 187), the author admits that both the bias toward Canada and the presence of ghosts of 1930 could be indications that politicians responded to the wishes of their voting public in an effort to secure reelection. It is accepted by many students of electoral processes that elections cannot be guaranteed to maximize social welfare; Lavergne tends to side with those others who see pressure group politics merely as a method of providing "weighted" votes, by which a rough kind of Kaldor-Scitovsky compensation rule is introduced into democratic electoral processes (pp. 11–20, esp. p. 18). The amount of print he devotes to investigating a situation in which a politician can have his cake (maximize social welfare) and eat it too (secure reelection) lends some support to this reviewer's belief that Lavergne thinks the American political and bureaucratic system produces, in general, economically efficient public decisions.

These findings (strong echoes of 1930; bias toward Canada) do not admit the striking interpretation that public behavior differs qualitatively from private, in that in the first but not in the second moral responsibility dominates an economistic calculus. Nonprice variables such as income, age, sex, and, if relevant, indices (such as his religion) of the moral attitudes of the demander are included in ordinary statistical demand analysis to ensure that estimates are not biased. To find that income is a very significant and large determinant of demand is not to deny that consumers economize (i.e., is not to deny the first law of demand); similarly, a finding of a large, significant coefficient for a moral variable does not indicate that politicians do not economize. The real test is whether they react to the relative costs of alternatives or not. In an ordinary consumer market, prices are relatively easy to observe; for the "markets" where politicians and bureaucrats operate, not so easy. We must therefore rely on some price proxies. The share of inputs sourced from Canada could be such a proxy.

In his rejection of the economistic model, Lavergne relies more on the lack of significance of what he calls "pressure group variables" than on the significance of the variables reflecting "principled behaviour." For it is against the "pressure group" explanation of tariffs that his fire is mostly directed, I believe, and not against the claim that people behave in the same way in politics as they do in the market. (It may seem unfair for a reviewer to pick up on a footnote at this point, but the one appended by Lavergne to the words quoted above from p. 12 is important, in that he there resiles from the claim of a qualitative difference by putting it merely as "one of degree," i.e., quantitative.) My own reading of the empirical literature is that there has been too great a concentration on pressure group demand for tariffs, and too little on supply. It is therefore possibly understandable that Lavergne's straw man is the advocate of a demand-dominated pressure group model, armed with one blade of the scissors.

Included in his class of pressure group variables is, for example, industrial concentration. Following Mancur Olson (1965), we all realize that a highly concentrated industry will suffer less than will an unconcentrated industry from free riding in the operation of a pressure group, but a casual observer of American politics would also suspect that government handouts to a highly concentrated industry (e.g., one with a "dominant firm") might evoke particularly vigorous public criticism, or a government rollback elicit public support (e.g., President Kennedy's intervention over steel prices). Similarly, if employment is geographically dispersed, employees are unlikely to form an active direct pressure group, and yet a politician seeking national office, or one wishing to implement a set of national policy goals that require the cooperation of the Senate, might do well not to damage the industry (take, for example, textiles in the United States). From a single-equation model it is impossible to tell whether a variable such as industrial concentration is all-around unimportant or whether two large effects, one on supply and one on demand, have happened to cancel each other out, a finding of great historical interest, if true.

Our confidence in its truth is, und [illegible]
made by Dougan (1984), using [illegible]
([1976]; not in Lavergne's bib [illegible]
Dougan shows that a reduced- [illegible]
variables as the industrial conc [illegible]
sumption, and the elasticity of [illegible]
gues for a positive coefficient for [illegible]
levels: the larger the share, the [illegible]
shows that the sign of this share can be [illegible]
relative political powers of the protected [illegible]
put, and the seekers of tariff revenue. To impose [illegible]
industry in the sample is, therefore, potentially a serious misspecification.

To illustrate the same kind of problem in the context of explanations of the fall in tariff rates, consider the effect of the introduction of federal income taxes in 1913, which can be taken to have reduced the force (exerted by revenue seekers) pushing rates of duty toward their revenue-maximizing levels. Freed substantially from their revenue-raising role, import duties would be pushed up if they were above revenue-maximizing levels, or pushed down if below. That is, the effect on tariffs of the income tax amendment varied according to whether the balance of political power had previously brought forth a duty that put import demand within the elastic or within the inelastic segment. A linear regression model is inappropriate.

It could well be true that pressure group variables or, more generally, the "economistic model" has been irrelevant to the levels of or changes in American tariffs since 1930. Nonetheless, Lavergne's book does not securely establish this. What it does do, and do well, is to draw on the political, historical, and institutional context for a cross-sectional tariff model.

JONATHAN J. PINCUS

Monash University
Australian National University

References

Dougan, William R. "Tariffs and the Economic Theory of Regulation." *Res. Law and Econ.*, vol. 6 (1984), in press

Hillman, Arye L. "Declining Industries and Political-Support Protectionist Motives." *A.E.R.* 72 (December 1982): 1180–87.

Olson, Mancur. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, Mass.: Harvard Univ. Press, 1965.

Peltzman, Sam. "Toward a More General Theory of Regulation." *J. Law and Econ.* 19 (August 1976). 211–48.

Journal of
Political Economy

EDITED BY ROBERT J. BARRO, JAMES J. HECKMAN,
JOSÉ A. SCHEINKMAN, AND GEORGE J. STIGLER
IN COOPERATION WITH OTHER MEMBERS OF THE
DEPARTMENT OF ECONOMICS AND THE
GRADUATE SCHOOL OF BUSINESS OF
THE UNIVERSITY OF CHICAGO
AND OUTSIDE REFEREES

Volume 92

February–December 1984

THE UNIVERSITY OF CHICAGO PRESS
CHICAGO, ILLINOIS

© 1984 BY THE UNIVERSITY OF CHICAGO. ALL RIGHTS RESERVED
PUBLISHED FEBRUARY,
APRIL, JUNE, AUGUST, OCTOBER,
DECEMBER 1984
BY THE UNIVERSITY OF CHICAGO PRESS

Index to Volume 92

Articles and Notes

Albrecht, James W., and Axell, Bo. An Equilibrium Model of Search Unemployment 824
Alston, Lee J.; Datta, Samar K., and Nugent, Jeffrey B. Tenancy Choice in a Competitive Framework with Transactions Costs 1121
Axell, Bo. See *Albrecht, James W*
Barnett, William A.; Offenbacher, Edward K., and Spindt, Paul A. The New Divisia Monetary Aggregates 1049
Baumol, William J., and Wolff, Edward N. On Interindustry Differences in Absolute Productivity 1017
Boulier, Bryan L., and Rosenzweig, Mark R. Schooling, Search, and Spouse Selection: Testing Economic Theories of Marriage and Household Behavior 712
Brecher, Richard A., and Choudhri, Ehsan U. New Products and the Factor Content of International Trade 965
Browning, Edgar K., and Johnson, William R. The Trade-Off between Equality and Efficiency 175
Bulow, Jeremy I., and Summers, Lawrence H. The Taxation of Risky Assets 20
Burmeister, Edwin. Sraffa, Labor Theories of Value, and the Economics of Real Wage Rate Determination 508
Caves, Richard E. Harry Johnson as a Social Scientist 642
Choudhri, Ehsan U. See *Brecher, Richard A*
Clark, Truman A. Violations of the Gold Points, 1890–1908 791
Corden, W. M. Harry Johnson's Contribu Theory
Datta, Samar K. See *Alston, Lee J*
Daughety, Andrew F. Regulation a
Dooley, Martin D., and Gottschalk, Pe in the United States: Trends Growth
Du Rietz, Gunnar. See *Hause, John*
Eckstein, Zvi. A Rational Expectatio
Evans, Paul. The Effects on Output of Volatility in the United States
Fane, George. The Incidence of a Tax on Pure Rent: The Old Reason for the Old Answer 329
Farzin, Y. Hossein. The Effect of the Discount Rate on Depletion of Exhaustible Resources 841
Flood, Robert P., and Garber, Peter M. Gold Monetization and Gold Discipline 90
Friedman, Daniel; Harrison, Glenn W.; and Salmon, Jon W. The Informational Efficiency of Experimental Asset Markets 349
Friedman, Milton. Financial Futures Markets and Tabular Standards 165
Garber, Peter M. See *Flood, Robert P*

Gift, Richard, and Marxsen, William. Aggregation and the Factoral Content of Trade 979
Gisser, Micha. Price Leadership and Dynamic Aspects of Oligopoly in U.S. Manufacturing 1035
Gottschalk, Peter. See *Dooley, Martin D.*
Grossman, Gene M., and Razin, Assaf. International Capital Movements under Uncertainty 286
Halvorsen, Robert, and Smith, Tim R. On Measuring Natural Resource Scarcity 954
Harberger, Arnold C., and Wall, David. Harry G. Johnson as a Development Economist 616
Harris, Milton, and Weiss, Yoram. Job Matching with Finite Horizon and Risk Aversion 758
Harrison, Glenn W. See *Friedman, Daniel.*
Hause, John C., and Du Rietz, Gunnar. Entry, Industry Growth, and the Microdynamics of Industry Supply 733
Haveman, Robert H., and Wolfe, Barbara L. The Decline in Male Labor Force Participation: Comment 532
Helpman, Elhanan. A Simple Theory of International Trade with Multinational Corporations 451
Johnson, William R. See *Browning, Edgar K.*
Jovanovic, Boyan. Matching, Turnover, and Unemployment 108
Kihlstrom, Richard E., and Riordan, Michael H. Advertising as a Signal 427
Kimbrough, Kent P. Aggregate Information and the Role of Monetary Policy in an Open Economy 268
Kimmel, Sheldon. A Note on Extraction with Nonconvex Costs 1158
Kormendi, Roger C., and Meguire, Philip G. Cross-Regime Evidence of Macroeconomic Rationality 875
Laidler, David. Harry Johnson as a Macroeconomist 592
Laidler, David. Misconceptions about the Real-Bills Doctrine: A Comment on Sargent and Wallace 149
Lee, Joong-Koon, and Wellington, Donald C. Angell and the Stable Money Rule 972
Longawa, Vicky M., editor and compiler. Harry G. Johnson: A Bibliography 659
McCallum, Bennett T. Are Bond-financed Deficits Inflationary? A Ricardian Analysis 123
McCormick, Robert E., and Tollison, Robert D. Crime on the Court 223
Malcomson, James M. Work Incentives, Hierarchy, and Internal Labor Markets 486
Marcus, Alan J. Efficient Asset Portfolios and the Theory of Normal Backwardation: A Comment 162
Marcus, Alan J., and Modest, David M. Futures Markets and Production Decisions 409
Marxsen, William. See *Gift, Richard.*
Meguire, Philip G. See *Kormendi, Roger C.*
Miller, Robert A. Job Matching and Occupational Choice 1086
Miyazaki, Hajime. On Success and Dissolution of the Labor-managed Firm in the Capitalist Economy 909
Modest, David M. See *Marcus, Alan J.*
Mumy, Gene E. Competitive Equilibria in Exhaustible Resource Markets with Decreasing Costs: A Comment on Eswaran, Lewis, and Heaps's Demonstration of Nonexistence 1168

Neftci, Salih N. Are Economic Time Series Asymmetric over the Business Cycle? 307
Nerlove, Marc, Razin, Assaf; and Sadka, Efraim Bequests and the Size of Population When Population Is Endogenous 527
Novos, Ian E., and Waldman, Michael. The Effects of Increased Copyright Protection: An Analytic Approach 236
Nugent, Jeffrey B. See *Alston, Lee J.*
Offenbacher, Edward K. See *Barnett, William A.*
Parsons, Donald O. Disability Insurance and Male Labor Force Participation: A Response to Haveman and Wolfe 542
Peters, Michael. Restrictions on Price Advertising 472
Razin, Assaf. See *Grossman, Gene M.*
Razin, Assaf. See *Nerlove, Marc.*
Riordan, Michael H. See *Kihlstrom, Richard E.*
Roberts, Russell D. A Positive Model of Private Charity and Public Transfers 136
Rosenzweig, Mark R. See *Boulier, Bryan L.*
Rotemberg, Julio J. A Monetary Equilibrium Model with Transactions Costs 40
Roth, Alvin E. The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory 991
Ruback, Richard S., and Zimmerman, Martin B. Unionization and Profitability: Evidence from the Capital Market 1134
Sadka, Efraim See *Nerlove, Marc*
Salmon, Jon W. See *Friedman, Daniel*
Seidman, Laurence S. Conversion to a Consumption Tax: The Transition in a Life-Cycle Growth Model 247
Shefrin, Hersh M. Inferior Forecasters, Cycles, and the Efficient-Markets Hypothesis: A Comment 156
Smith, Tim R. See *Halvorsen, Robert*
Spindt, Paul A. See *Barnett, William A.*
Summers, Lawrence H. See *Bulow, Jeremy I.*
Tollison, Robert D. See *McCormick, Robert E.*
Waldman, Michael See *Novos, Ian E.*
Wall, David See *Harberger, Arnold C.*
Weiss, Yoram See *Harris, Milton*
Wellington, Donald C. See *Lee, Joong*
Wolfe, Barbara L. See *Haveman, R*
Wolff, Edward N. See *Baumol, Will*
Wolpin, Kenneth I. An Estimable D and Child Mortality ...
Zimmerman, Martin B. See *Ruback, Richard S.*

Reviews

Allais, Maurice, *La Théorie générale des surplus.* By Roger Guesnerie 780
Bhagwati, Jagdish N., ed., *Import Competition and Response.* By Rachel McCulloch 550
Chick, Victoria, *Macroeconomics after Keynes: A Reconsideration of the General Theory.* By Robert M. Solow 784
Edelstein, Michael, *Overseas Investment in the Age of High Imperialism: The United Kingdom, 1850–1914.* By C. Knick Harley 334

Friedman, Joseph, and Weinberg, Daniel H., *The Economics of Housing Vouchers.* By John C. Weicher 985
Frisch, Helmut, ed., *Schumpeterian Economics.* By Scott Gordon 168
Fuchs, Victor R., *How We Live: An Economic Perspective on Americans from Birth to Death.* By Lee Benham 787
Gossen, Herman H., *The Laws of Human Relations.* By George J. Stigler 1175
Hahn, Frank, *Money and Inflation.* By Herschel I. Grossman 337
Hall, Robert E., ed., *Inflation: Causes and Effects.* By Jeremy Greenwood 553
Hall, Robert E., and Rabushka, Alvin, *Low Tax, Simple Tax, Flat Tax.* By Joseph A. Pechman 340
Kalt, Joseph P., *The Economics and Politics of Oil Price Regulation: Federal Policy in the Post Embargo Era.* By Henry D. Jacoby 1176
Lavergne, Réal P., *The Political Economy of U.S. Tariffs: An Empirical Analysis.* By Jonathan J. Pincus 1178
Mueller, Dennis C., ed., *The Political Economy of Growth.* By Anne O. Krueger 558
Nicholson, T. R., *The Birth of the British Motor Car, 1769–1897.* By Nathan Rosenberg 988
Nove, Alec, *The Economics of Feasible Socialism* By Nicholas Spulber .. 170
Rosenn, Keith S., *Law and Inflation.* By Meir Heth 343
Sattinger, Michael, *Capital and the Distribution of Labor Earnings.* By Alan Harrison 560
Schultz, T. Paul, *Economics of Population.* By Boone A. Turchi 345